# Best of LessWrong: December 2017

1. Goodhart Taxonomy
2. Updates from Boston
3. More Dakka
4. In the presence of disinformation, collective epistemology requires local modeling
5. Against Love Languages
6. Show LW: Diaspora Project Map and Preregistration Database
7. Cash transfers are not necessarily wealth transfers
8. Mapping the Social Mind (Buttons)
9. Rules of variety
10. TSR #7: Universal Principles
11. Epistemic Spot Check: Full Catastrophe Living (Jon Kabat-Zinn)
12. Philosophy of Numbers (part 2)
13. Comment on SSC's Review of Inadequate Equilibria
14. Thinking as the Crow Flies: Part 2 - Basic Logic via Precommitments
15. Empirical philosophy and inversions
16. Oracle paper
17. Calling Bullshit - Lectures and Readings on Evaluating Scientific Research
18. 2017 AI Safety Literature Review and Charity Comparison
19. The Basic Object Model and Definition by Interface
20. Truth is Symmetric
21. Learning AI if you suck at math
22. Melting Gold, and Organizational Capacity
23. Against the Linear Utility Hypothesis and the Leverage Penalty
24. Thinking as the Crow Flies: Part 3 - Tokens, Syntax, and Expressions
25. Writing Down Conversations
26. TSR #5 The Nature of Operations
27. Improvement Without Superstition
28. The map of "Levels of defence" in AI safety
29. Guarding Slack vs Substance
30. Success and Fail Rates of Monthly Policies
31. Towards a Rigorous Model of Virtue-Signalling
32. Pascal's Muggle Pays
33. Can we see light?
34. A List Of Questions & Exercises For Reviewing Your Year
35. Why did everything take so long?
36. TSR #6: Strength and Weakness
37. Quick thoughts on empathic metaethics
38. Maps vs Buttons; Nerds vs Normies
39. Methods of Phenomenology
40. Happiness Is a Chore
41. Comments on Power Law Distribution of Individual Impact
42. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm
43. The World as Phenomena
44. Thinking as the Crow Flies: Part 1 - Introduction
45. Mana
46. Conceptual Similarity Does Not Imply Actionable Similarity
47. 12/12/2017 Update: Creating Sequences
48. The expected value of the long-term future
49. Philosophy of Numbers (part 1)
50. Bayes and Paradigm Shifts - or being wrong af

# Best of LessWrong: December 2017

1. Goodhart Taxonomy
2. Updates from Boston
3. More Dakka
4. In the presence of disinformation, collective epistemology requires local modeling
5. Against Love Languages
6. Show LW: Diaspora Project Map and Preregistration Database
7. Cash transfers are not necessarily wealth transfers
8. Mapping the Social Mind (Buttons)
9. Rules of variety
10. TSR #7: Universal Principles
11. Epistemic Spot Check: Full Catastrophe Living (Jon Kabat-Zinn)
12. Philosophy of Numbers (part 2)
13. Comment on SSC's Review of Inadequate Equilibria
14. Thinking as the Crow Flies: Part 2 - Basic Logic via Precommitments
15. Empirical philosophy and inversions
16. Oracle paper
17. Calling Bullshit - Lectures and Readings on Evaluating Scientific Research
18. 2017 AI Safety Literature Review and Charity Comparison
19. The Basic Object Model and Definition by Interface
20. Truth is Symmetric
21. Learning AI if you suck at math
22. Melting Gold, and Organizational Capacity
23. Against the Linear Utility Hypothesis and the Leverage Penalty
24. Thinking as the Crow Flies: Part 3 - Tokens, Syntax, and Expressions
25. Writing Down Conversations
26. TSR #5 The Nature of Operations
27. Improvement Without Superstition
28. The map of "Levels of defence" in AI safety
29. Guarding Slack vs Substance
30. Success and Fail Rates of Monthly Policies
31. Towards a Rigorous Model of Virtue-Signalling
32. Pascal’s Muggle Pays
33. Can we see light?
34. A List Of Questions & Exercises For Reviewing Your Year
35. Why did everything take so long?
36. TSR #6: Strength and Weakness
37. Quick thoughts on empathic metaethics
38. Maps vs Buttons; Nerds vs Normies
39. Methods of Phenomenology
40. Happiness Is a Chore
41. Comments on Power Law Distribution of Individual Impact
42. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm
43. The World as Phenomena
44. Thinking as the Crow Flies: Part 1 - Introduction
45. Mana
46. Conceptual Similarity Does Not Imply Actionable Similarity
47. 12/12/2017 Update: Creating Sequences

48. The expected value of the long-term future
49. Philosophy of Numbers (part 1)
50. Bayes and Paradigm Shifts - or being wrong af

# Goodhart Taxonomy

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

Goodhart's Law states that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." However, this is not a single phenomenon. I propose that there are (at least) four different mechanisms through which proxy measures break when you optimize for them.

The four types are Regressional, Causal, Extremal, and Adversarial. In this post, I will go into detail about these four different Goodhart effects using mathematical abstractions as well as examples involving humans and/or AI. I will also talk about how you can mitigate each effect.

Throughout the post, I will use $V$ to refer to the true goal and use $U$ to refer to a proxy for that goal which was observed to correlate with $V$ and which is being optimized in some way.

---

# Quick Reference

- Regressional Goodhart - When selecting for a proxy measure, you select not only for the true goal, but also for the difference between the proxy and the goal.
  - Model: When $U$ is equal to $V + X$, where $X$ is some noise, a point with a large $U$ value will likely have a large $V$ value, but also a large $X$ value. Thus, when $U$ is large, you can expect $V$ to be predictably smaller than $U$.
  - *Example: height is correlated with basketball ability, and does actually directly help, but the best player is only 6'3", and a random 7' person in their 20s would probably not be as good*
- Causal Goodhart - When there is a non-causal correlation between the proxy and the goal, intervening on the proxy may fail to intervene on the goal.
  - Model: If $V$ causes $U$ (or if $V$ and $U$ are both caused by some third thing), then a correlation between $V$ and $U$ may be observed. However, when you intervene to increase $U$ through some mechanism that does not involve $V$, you will fail to also increase $V$.
  - *Example: someone who wishes to be taller might observe that height is correlated with basketball skill and decide to start practicing basketball.*
- Extremal Goodhart - Worlds in which the proxy takes an extreme value may be very different from the ordinary worlds in which the correlation between the proxy and the goal was observed.
  - Model: Patterns tend to break at simple joints. One simple subset of worlds is those worlds in which $U$ is very large. Thus, a strong correlation between

U and V observed for naturally occuring U values may not transfer to worlds in which U is very large. Further, since there may be relatively few naturally occuring worlds in which U is very large, extremely large U may coincide with small V values without breaking the statistical correlation.
    - *Example: the tallest person on record, Robert Wadlow, was 8'11" (2.72m). He grew to that height because of a pituitary disorder, he would have struggled to play basketball because he "required leg braces to walk and had little feeling in his legs and feet."*
- Adversarial Goodhart - When you optimize for a proxy, you provide an incentive for adversaries to correlate their goal with your proxy, thus destroying the correlation with your goal.
    - Model: Consider an agent A with some different goal W. Since they depend on common resources, W and V are naturally opposed. If you optimize U as a proxy for V, and A knows this, A is incentivized to make large U values coincide with large W values, thus stopping them from coinciding with large V values.
    - *Example: aspiring NBA players might just lie about their height.*

---

# Regressional Goodhart

When selecting for a proxy measure, you select not only for the true goal, but also for the difference between the proxy and the goal.

## Abstract Model

When U is equal to V + X, where X is some noise, a point with a large U value will likely have a large V value, but also a large X value. Thus, when U is large, you can expect V to be predictably smaller than U.

The above description is when U is meant to be an estimate of V. A similar effect can be seen when U is only meant to be correlated with V by looking at percentiles. When a sample is chosen which is a typical member of the top p percent of all U values, it will have a lower V value than a typical member of the top p percent of all V values. As a special case, when you select the highest U value, you will often not select the highest V value.

## Examples

Examples of Regressional Goodhart are everywhere. Every time someone does something that is anything other than the thing that maximizes their goal, you could view it as them optimizing some kind of proxy (and the action to maximize the proxy is not the same as the action to maximize the goal).

Regression to the Mean, Winner's Curse, and Optimizer's Curse are all examples of Regressional Goodhart, as is the Tails Come Apart phenomenon.

## Relationship with Other Goodhart Phenomena

Regressional Goodhart is by far the most benign of the four Goodhart effects. It is also the hardest to avoid, as it shows up every time the proxy and the goal are not exactly the same.

## Mitigation

When facing only Regressional Goodhart, you still want to choose the option with the largest proxy value. While the proxy will be an overestimate it will still be better in expectation than options with a smaller proxy value. If you have control over what proxies to use, you can mitigate Regressional Goodhart by choosing proxies that are more tightly correlated with your goal.

If you are not just trying to pick the best option, but also trying to have an accurate picture of what the true value will be, Regressional Goodhart may cause you to overestimate the value. If you know the exact relationship between the proxy and the goal, you can account for this by just calculating the expected goal value for a given proxy value. If you have access to a second proxy with an error independent from the error in the first proxy, you can use the first proxy to optimize, and the second proxy to get an accurate expectation of the true value. (This is what happens when you set aside some training data to use for testing.)

---

# Causal Goodhart

When there is a non-causal correlation between the proxy and the goal, intervening on the proxy may fail to intervene on the goal.

## Abstract Model

If $V$ causes $U$ (or if $V$ and $U$ are both caused by some third thing), then a correlation between $V$ and $U$ may be observed. However, when you intervene to increase $U$ through some mechanism that does not involve $V$, you will fail to also increase $V$.

## Examples

Humans often avoid naive Causal Goodhart errors, and most examples I can think of sound obnoxious (like eating caviar to become rich). One possible example is a human who avoids doctor visits because not being told about health is a proxy for being healthy. (I do not know enough about humans to know if Causal Goodhart is actually what is going on here.)

I also cannot think of a good AI example. Most AI is not in acting in the kind of environment where Causal Goodhart would be a problem, and when it is acting in that kind of environment Causal Goodhart errors are easily avoided.

Most of the time the phrase "Correlation does not imply causation" is used it is pointing out that a proposed policy might be subject to Causal Goodhart.

## Relationship with Other Goodhart Phenomena

You can tell the difference between Causal Goodhart and the other three types because Causal Goodhart goes away when just sample a world with large proxy value, rather than intervene to cause the proxy to happen.

## Mitigation

One way to avoid Causal Goodhart is to only sample from or choose between worlds according to their proxy values, rather than causing the proxy. This clearly cannot be done in all situations, but it is useful to note that there is a class of problems for which Causal Goodhart cannot cause problems. For example, consider choosing between algorithms based on how well they do on some test inputs, and your goal is to choose an algorithm that performs well on random inputs. The fact that you choose an algorithm does not effect its performance, and you don't have to worry about Causal Goodhart.

In cases where you actually change the proxy value, you can try to infer the causal structure of the variables using statistical methods, and check that the proxy actually causes the goal before you intervene on the proxy.

---

# Extremal Goodhart

Worlds in which the proxy takes an extreme value may be very different from the ordinary worlds in which the correlation between the proxy and the goal was observed.

## Abstract Model

Patterns tend to break at simple joints. One simple subset of worlds is those worlds in which $U$ is very large. Thus, a strong correlation between $U$ and $V$ observed for

naturally occuring $U$ values may not transfer to worlds in which $U$ is very large.

Further, since there may be relatively few naturally occuring worlds in which $U$ is very large, extremely large $U$ may coincide with small $V$ values without breaking the statistical correlation.

# Examples

Humans evolve to like sugars, because sugars were correlated in the ancestral environment (which has fewer sugars) with nutrition and survival. Humans then optimize for sugars, have way too much, and become less healthy.

As an abstract mathematical example, let $U$ and $V$ be two correlated dimensions in a multivariate normal distribution, but we cut off the normal distribution to only include the ball of points in which $U^2 + V^2 < n$ for some large n. This example represents a correlation between $U$ and $V$ in naturally occurring points, but also a boundary around what types of points are feasible that need not respect this correlation. Imagine you were to sample k points and take the one with the largest $U$ value. As you increase k, at first, this optimization pressure lets you find better and better points for both $U$ and $V$, but as you increase k to infinity, eventually you sample so many points that you will find a point near $U = n, V = 0$. When enough optimization pressure was applied, the correlation between $U$ and $V$ stopped mattering, and instead the boundary of what kinds of points were possible at all decided what kind of point was selected.

Many examples of machine learning algorithms doing bad because of overfitting are a special case of Extremal Goodhart.

# Relationship with Other Goodhart Phenomena

Extremal Goodhart differs from Regressional Goodhart in that Extremal Goodhart goes away in simple examples like correlated dimensions in a multivariate normal distribution, but Regressional Goodhart does not.

# Mitigation

Quantilization and Regularization are both useful for mitigating Extremal Goodhart effects. In general, Extremal Goodhart can be mitigated by choosing an option with a high proxy value, but not so high as to take you to a domain drastically different from the one in which the proxy was learned.

# Adversarial Goodhart

When you optimize for a proxy, you provide an incentive for adversaries to correlate their goal with your proxy, thus destroying the correlation with your goal.

## Abstract Model

Consider an agent A with some different goal W. Since they depend on common resources, W and V are naturally opposed. If you optimize U as a proxy for V, and A knows this, A is incentivized to make large U values coincide with large W values, thus stopping them from coinciding with large V values.

## Examples

When you use a metric to choose between people, but then those people learn what metric you use and game that metric, this is an example of Adversarial Goodhart.

Adversarial Goodhart is the mechanism behind a superintelligent AI making a Treacherous Turn. Here, V is doing what the humans want forever. U is doing what the humans want in the training cases where the AI does not have enough power to take over, and W is whatever the AI wants to do with the universe.

Adversarial Goodhart is also behind the malignancy of the universal prior, where you want to predict well forever (V), so hypotheses might predict well for a while (U), so that they can manipulate the world with their future predictions (W).

## Relationship with Other Goodhart Phenomena

Adversarial Goodhart is the primary mechanism behind the original Goodhart's Law.

Extremal Goodhart can happen even without any adversaries in the environment. However, Adversarial Goodhart may take advantage of Extremal Goodhart, as an adversary can more easily manipulate a small number of worlds with extreme proxy values, than it can manipulate all of the worlds.

## Mitigation

Succesfully avoiding Adversarial Goodhart problems is very difficult in theory, and we understand very little about how to do this. In the case of non-superintelligent adversaries, you may be able to avoid Adversarial Goodhart by keeping your proxies secret (for example, not telling your employees what metrics you are using to

evaluate them). However, this is unlikely to scale to dealing with superintelligent adversaries.

One technique that might help in mitigating Adversarial Goodhart is to choose a proxy that is so simple and optimize so hard that adversaries have no or minimal control over the world which maximizes that proxy. (I want to ephasize that this is not a good plan for avoiding Adversarial Goodhart; it is just all I have.)

For example, say you have a complicated goal that includes wanting to go to Mars. If you use a complicated search process to find a plan that is likely to get you to Mars, adversaries in your search process may suggest a plan that involves building a superintelligence that gets you to Mars, but also kills you.

On the other hand, if you use the proxy of getting to Mars as fast as possible and optimize very hard, then (maybe) adversaries can't add baggage to a proposed plan without being out selected by a plan without that baggage. Buliding a superintelligence maybe takes more time than just having the plan tell you how to build a rocket quickly. (Note that the plan will likely include things like acceleration that humans can't handle and nanobots that don't turn off, so Extremal Goodhart will still kill you.)

# Updates from Boston

There isn't enough sharing of positive and negative results within the rationality community[1]. I suspect this results in a fair amount of wasted effort as people explore the same dead ends, and a fair amount of lost potential when more effective tools don't get shared[2].

So, here are some things Boston has tried (not everything though):

# Successes

## Bureaucracy Day

Everyone shows up for a few hours with the intention of taking care of whatever bureaucratic tasks they've been putting off (doctor's appointments, getting a passport, taking care of personal finance tasks, etc.). Various supplies (printer, staplers, envelopes, etc.) are available.

I record how long each attempted task had been put off for, whether or not it was completed, and (optionally) what the task was. I use "old tasks get accomplished" as a proxy for the impact of the intervention, since I assume that if someone has been putting something off for six months they weren't going to do it in the counterfactual world where Bureaucracy Day doesn't exist.

Overall it's been *way* more successful than I expected. It's not uncommon for tasks that are *years* old to get finished.

I expect efficacy to drop over time as all the oldest tasks get accomplished, but so far that's been counteracted by new people participating.

## Sprint Day

My attempt to use Hackathon mindset for something actually productive. People show up and work for ten hours. No social media, no non-essential conversations, and only one project is allowed. Food is ordered or cooked the night before.

There aren't any common metrics, because I don't want to disrupt workflow by imposing recording procedures. Using my own metrics (number of github issues resolved and time spent working) I'm about an order of magnitude more productive on sprint days than on normal Saturdays. Not enough data to tell if I'm just redistributing my productivity to sprint days, but the data do not suggest this. Other participants report excellent results (that sprint days are at least 90th percentile productivity).

However participation has been *very* low (and I suspect this hurts individual efficacy). I'm not sure if this is because people don't want to give up their Saturdays, don't have a project to work on, or something else.

## Backthumb

A conversational norm that anyone can say "backthumb" and the conversation will return to the previous topic. We use it to cull tangents. It's useful enough to have reached fixation within the local community, and we use it regularly.

I do not have any data on how focused our conversations are with / without the backthumb norm, but I'm quite confident it's a significant improvement.

### Boiling point of nitrogen

Another conversational norm. Anyone can say "boiling point of nitrogen" to indicate that the current disagreement/question can be easily resolved with google, and then everyone has to shut up and google it. (Not sure if this is original to Boston or if we imported it).

Works well for culling pointless debates. Adoption has steadily increased.

## Failures

### Order of the Sphex

We made a weekly review worksheet, and attempted to iterate on it. Questions were things like: "What goals are you working on", "What trivial inconveniences are in your way", "Is there something you need to get off your plate". (I can share the full list if anyone is interested).

Was initially successful, but eventually became useless, and attempts to save it failed. There was also a meta failure where we didn't notice how badly it was failing, and so continued spending time on it.

### Timed worksheets

There were many attempts to make use of time worksheets (a la CFAR) in our early days. As far as I can tell these were almost never useful for anyone (although one person reports them being very useful).

### Group Habitica

Individual members have reported finding group Habiticas useful in the past, but our attempt to create a community-wide Habitica failed. Very few people joined, and of those very few made use of it.

### Group intervention testing

We created a list of interventions (take modafinil, plan your day in the morning, etc.). Once a week an intervention was picked at random, and everyone would try it.

Project died almost immediately, since no one actually implemented the chosen interventions.

[1] : Or people aren't trying new things (I hope not) or they're sharing but in places I don't check.

[2] I'm not claiming that all results will generalize, or that it's never worth replicating an idea. But currently those aren't even possibilities (subject to [1] ).

# More Dakka

Epistemic Status: Hopefully enough Dakka

Eliezer Yudkowsky's book Inadequate Eqilibria is excellent. I recommend reading it, if you haven't done so. Three recent reviews are Scott Aaronson's, Robin Hanson's (which inspired You Have the Right to Think and a great discussion in its comments) and Scott Alexander's. Alexander's review was an excellent summary of key points, but like many he found the last part of the book, ascribing much modesty to status and prescribing how to learn when to trust yourself, less convincing.

My posts, including Zeroing Out and Leaders of Men have been attempts to extend the last part, offering additional tools. Daniel Speyer offers good concrete suggestions as well. My hope here is to offer both another concrete path to finding such opportunities, and additional justification of the central role of social control (as opposed to object-level concerns) in many modest actions and modesty arguments.

Eliezer uses several examples of civilizational inadequacy. Two central examples are the failure of the Bank of Japan and later the European Central Bank to print sufficient amounts of money, and the failure of anyone to try treating seasonal affective disorder with sufficiently intense artificial light.

In a MetaMed case, a patient suffered from a disease with a well-known reliable biomarker and a safe treatment. In studies, the treatment improved the biomarker linearly with dosage. Studies observed that sick patients whose biomarkers reached healthy levels experienced full remission. The treatment was fully safe. No one tried increasing the dose enough to reduce the biomarker to healthy levels. If they did, they never reported their results.

In his excellent post Sunset at Noon, Raymond points out Gratitude Journals:

> *"Rationalists obviously don't *actually* take ideas seriously. Like, take the Gratitude Journal. This is the one peer-reviewed intervention that *actually increases your subjective well being*, and costs barely anything. And no one I know has even seriously tried it. Do literally *none* of these people care about their own happiness?"*

*"Huh. Do *you* keep a gratitude journal?"*

*"Lol. No, obviously."*

*– Some Guy at the Effective Altruism Summit of 2012*

Gratitude journals are awkward interventions, as Raymond found, and we need to find details that make it our own, or it won't work. But the active ingredient, *gratitude,* obviously works and is freely available. Remember the last time someone expressed gratitude to you and it made your day *worse?* Remember the last time you expressed gratitude to someone else, or felt gratitude about someone or something, and it made *your* day worse?

In my experience it happens *approximately zero times.* Gratitude just works, unmistakably. I once sent a single gratitude letter. It increased my baseline well-being. Then I didn't write more. I do try to remember to feel gratitude, and express it.

That helps. But I can't think of a good reason *not* to do that *more*, or for *anyone* I know to not do it more.

In all four cases, our civilization has (it seems) correctly found the solution. We've tested it. It works. The more you do, the better it works. There's probably a level where side effects would happen, but there's no sign of them yet.

We know the solution. Our bullets work. We just need more. We need More (and better) (metaphorical) Dakka. And then we decide we're out of bullets. We stop.

If it helps but doesn't solve your problem, *perhaps you're not using enough.*

I

We don't use enough to find out how much enough would be, or what bad things it might cause. More Dakka might backfire. It also might solve your problem.

The Bank of Japan didn't have enough money. They printed some. It helped a little. They could have kept printing more money until printing more money either solves their problem or starts to cause other problems. They didn't.

Yes, some countries printed too much money and very bad things happened, but no countries printed too much money *because they wanted more inflation.* That's not a thing.

Doctors saw patients suffer for lack of light. They gave them light. It helped a little. They could have tried more light until it solved their problem or started causing other problems. They didn't.

Yes,people suffer from too much sunlight, or spending too long in tanning beds, but those are skin conditions (as far as I know) and we don't have examples of too much of this kind of artificial light, other than it being unpleasant.

Doctors saw patients suffer from a disease in direct proportion to a biomarker. They gave them a drug. It helped a little, with few if any side effects. They could have increased the dose until it either solved the problem or started causing other problems. They didn't.

Yes, drug overdoses cause bad side effects, but we could find no record of *this* drug causing any bad side effects at any reasonable dosage, or any theory why it would.

People express gratitude. We are told it improves subjective well-being in studies. Our subjective well-being improves a little. We could express more gratitude, with no real downsides. Almost all of us don't.

On that note, thanks for reading!

A decision was universally made that enough, *despite obviously not being enough,* was enough. 'More' was never tried.

This is important on two levels.

II

The first level is practical. If you think a problem could be solved or a situation improved by More Dakka, there's a good chance you're right.

Sometimes a little more is a little better. Sometimes a lot more is a *lot* better. Sometimes each attempt is unlikely to work, but improves your chances.

If something is a good idea, *you need a reason* to not try doing more of it.

No, seriously. You need a reason.

The second level is, ‘do more of what is already working and see if it works more’ is as basic as it gets. If we can’t reliably try *that,* we can’t reliably try *anything.* How could you ever say ‘If that worked someone would have tried it’?

You can’t. If no one says they tried it, probably no one tried it. There might be *good reasons* not to try it. There also might not. There’d still be a good chance no one tried it.

There’s also a chance someone *did* try it and isn’t reporting the results anywhere you can find. That doesn’t mean it didn’t work, let alone that it can never work.

III

Why would this be an overlooked strategy?

It sounds crazy that it could be overlooked. It’s overlooked.

Eliezer gives three tools to recognize places systems fail, using highly useful economic arguments I recommend using frequently:

1. Cases where the decision lies in the hands of people who would gain little personally, or lose out personally, if they did what was necessary to help someone else;

2. Cases where decision-makers can’t reliably learn the information they need to make decisions, even though someone else has that information

3. Systems that are broken in multiple places so that no one actor can make them better, even though, in principle, some magically coordinated action could move to a new stable state.

In these cases, I do not think such explanations are enough.

If the Bank of Japan didn’t print more money, that implies the Bank of Japan wasn’t sufficiently incentivized to hit their inflation target. They must have been maximizing primarily for prestige instead. I can buy that, but why didn’t they think the best way to do that was to *hit the inflation target?* Alexander’s suggested payoff matrix, where printing more money makes failure much worse, isn’t good enough. It can’t be central on its own. The answer was too clear, the payoff worth the odds, and they had the information, as I detail later.

Eliezer gives the model of researchers looking for citations plus grant givers looking for prestige, as the explanation for why his SAD treatment wasn’t tested. I don’t buy it. Story doesn’t make sense.

If more light worked, you’d get a *lot* of citations, for not much cost or effort. If you’re writing a grant, this costs little money and could help many people.
It’s *less* prestigious to up the dosage than be original, but it’s still a big prestige win.

If you say they want to associate with high status research folk, then they won't care about the grant contents, so it reduces to a one-factor market, where again researchers should try this.

Alexander noticed the same confusion on that one.

In the drug dosage case, Eliezer's tools do better. No doctor takes the risk of being sued if something goes wrong, and no company makes money by funding the study and it's too expensive for a grant, and trying it on your own feels too risky. Maybe. It still does not feel like enough. The paths forward are too easy, too cheap, the payoff too large and obvious. Even one wealthy patient could break through, and it would be worth it. Yet even our patient, as far as we know, didn't even try it and certainly didn't report back.

The gratitude case doesn't fit the three modes at all.

IV

Here is my model.I hope it illuminates when to try such things yourself.

Two key insights here are The Thing and the Symbolic Representation of The Thing, and Scott Alexander's Concept-Shaped Holes Can Be Impossible To Notice. Both are worth reading, in that order.

I'll summarize the relevant points.

The standard amount of something, by definition, counts as the symbolic representation of the thing. The Bank of Japan 'printed money.' The standard SAD treatment 'exposes people to light.' Our patients' doctors prescribed 'standard drug.' Today, various people 'left with plenty of time,' 'came up with a plan,' 'were part of a community,' 'ate pizza,' 'listened to the other person,' 'focused on their breath,' 'bought enough nipple tops for the baby's bottles,' 'did their job' and 'added salt and pepper.'

They got results. A little. Better than nothing. But much less than was desired.

The Bank of Australia printed enough money. Eliezer Yudkowsky exposed his wife to enough light. Our patient was told to take enough of the drug to actually work. Meanwhile, other people actually left with plenty of time, actually came up with a workable plan, actually were part of a community, ate real pizza, actually listened to another person, actually focused on their breath, bought enough nipple tops for the baby's bottles, actually did their job, and added copious amounts of sea salt and freshly ground pepper.

Some of these are about *quality* rather than *quantity.* You could also think of that as a bigger quantity of effort, or willingness to pay more money or devote more time. Still, it's worth noting that an important variant of 'use more,' 'do more' or 'do more often' is 'do it better.'

Being part of that second group is harder than it looks:

You need to realize the thing might exist at all.

You need to realize the symbolic representation of the thing isn't the thing.

You need to ignore the idea that you've done your job.

You need to actually care about solving the problem.

You need to think about the problem a little.

You need to ignore the idea that no one could blame you for not trying.

You need to not care that what you're about to do is unusual or weird or socially awkward.

You need to not care that what you're about to do might be high status.

You need to not care that what you're about to do might be low status.

You need to not care that what you're about to do might not work.

You need to not be concerned that what you're about to do might work.

You need to not care that what you're about to do might backfire.

You need to not care that what you're about to do is immodest.

You need to not instinctively assume that this will backfire *because* attempting it would be immodest, so the world will find some way to strike you down.

You need to not care about the implicit accusation you're making against everyone who didn't try it.

You need to not care that what you're about to do might be wasteful. Or inappropriate. Or weird. Or unfair. Or morally wrong. Or something.

Why is this list getting so long? What is that answer of 'don't do it' doing on the bottom of the page?

V

Long list is long. A lot of items are related. Some will be obvious, some won't be. Let's go through the list.

**You need to realize the thing might exist at all.**

One cannot do better unless one realizes it might be possible to do better. Scott gives several examples of situations in which he doubted the existence of the thing.

**You need to realize the symbolic representation of the thing isn't the thing.**

Scott gives several examples where he thought he knew what the thing was, only to find out he had no idea; what he thought was the thing was actually a symbolic representation, a pale shadow. If you think having a few friends is what a community is, it won't occur to you to seek out a real one.

**You need to ignore the idea that you've done your job.**

There was a box marked 'thing'. You've checked that box off by getting the symbolic version of the thing. It's easy to then think you've done the job and are somehow

done. Even if you're doing this for yourself or someone you care about, there's this urge to get to and think 'job done', 'quest complete', and not think about details. You need to realize you're not doing the job so you can *say* you've done the job, or so you can *tell yourself* you've done the job. Even if you didn't get what you wanted, your real job was to get the right to tell a story you can tell yourself that you tried to get it, right?

**You need to actually care about solving the problem.**

You're doing the job *so the job gets done.* That's why doing the symbolic version doesn't mean you're done. Often people don't care much about solving the problem. They care whether they're responsible. They care whether socially appropriate steps have been taken.

**You need to ignore the idea that no one could blame you for not trying.**

Alexander notes how important this one is, and it's really big.

People often care primarily about doing that which no one could blame them for. Being blamed or scapegoated is really bad. Even self-blame! We instinctively fear someone will discover and expose us, and make ourselves feel bad. We cover up the evidence and create justifications. Doing the normal means no one could blame you. If you don't grasp that this is a thing, read as much of Atlas Shrugged as needed until you grasp that. It should only take a chapter or two, but *this idea alone is worth a thousand page book in order to get, if that's what it takes.* I'm not kidding.

Blame does happen. The real incentive here is big. The incentive people *think* they have to do this, even when the chance of being blamed is minimal, is *much, much* bigger.

**You need to think about the problem a little.**

People don't like thinking.

**You need to not care that what you're about to do is unusual or weird or socially awkward.**

There's a primal fear of doing anything unusual or weird. More would be unusual and weird. It might be slightly socially awkward. You'd never know until it actually *was* awkward. That would be *just awful.* Can't have that. No one is watching or cares, but some day someone might *find you out* and then expose you as no good. We go around being normal, only guessing which slightly weird things would get us in trouble, or that we'd need to get someone else in trouble for! So we try to do none of them. That's what happens when not operating on object-level causal models full of gears about what will work.

**You need to not care that what you're about to do might be high status.**

Doing or tying to do something high status is to claim high status. Claiming status you're not entitled to is a good way to get into a lot of trouble. Claiming to usefully think, or to know something, is automatically high status. Are you sure you have that right?

**You need to not care that what you're about to do might be low status.**

Your status would go down. That's even worse. If it's high status you lose, if it's low status you also lose, and you don't even know which one it is since no one does it! Might even be both. Better to leave the whole thing alone.

**You need to not care that what you're about to do might not work.**

Failing is *just awful.* Even things that are *supposed* to mostly fail. Even getting ludicrous odds. Only explicitly permitted narrow exceptions are permitted, which shrink each year. Otherwise we must, must succeed, or nothing we do will ever work and everyone will know that. I founded a company once*. It didn't work.
Now *everyone knows* rationalists can't found companies. Shouldn't have tried.

* – Well, three times.

**You need to not be concerned that what you're about to do might work.**

Even worse, it might work. Then what? No idea. Does not compute. You'd have to keep doing weird thing, or advocate for weird thing. How weird would that be? What about the people you'd prove wrong? What would you even say?

**You need to not care that what you're about to do might backfire.**

It might not only not work, it might have real consequences. That's a thing. Can't think of why that might happen. Every brainstormed risk seems highly improbable and not that big a deal. But why take that risk?

**You need to not care that what you're about to do is immodest.**

By modesty, anything you think of, that's worth thinking, has been thought of. Anything worth trying has been tried, anything worth doing done. Ignore that there's a first time for everything. Who are you to claim there's something worth trying? Who are you to claim you know better than everyone else? Did you not notice all the other people? Are you really high status enough to claim you know better than all of them? Let's see that hero licence of yours, buster. Object-level claims are status claims!

**You need to not instinctively assume that this will backfire *because* attempting it would be immodest, so the world will find some way to strike you down.**

The world won't let you *get away with that.* It will make this blow up in your face. And laugh. At you. People know this. They'll  instinctively join the conspiracy making it happen, coordinating seamlessly. Their alternative is thinking for themselves, or other people might thinking for themselves rather than playing imitation games. Unthinkable. Let's scapegoat someone and reinforce norms.

**You need to not care about the implicit accusation you're making against everyone who didn't try it.**

You're not only calling them *wrong*. You're saying *the answer was in front of their face the whole time.* They had an obvious solution and didn't take it. You're telling them they didn't have a good reason for that. They gonna be pissed.

**You need to not care that what you're about to do might be wasteful. Or inappropriate. Or unfair. Or low status. Or lack prestige. Or be morally wrong. Or something. There's gotta be something!**

The answer is right there at the bottom of the page. This isn't done, so don't do it. Find a reason. If there isn't a good one, go with what you got. Flail around as needed.

That's what the Bank of Japan was *actually* afraid of. Nothing. A vague feeling they were *supposed* to be afraid of *something*, so they kept brainstorming until something sounded plausible.

Printing money might mean printing too much! The opposite is true. Not printing money now means having to print even more later, as the economy suffers.

Printing money would destroy their credibility! The opposite is true. *Not* printing money destroyed their credibility.

People don't like it when we print too much money! The opposite is true. Everyone was yelling at them to print more money.

The markets don't like it when we print too much money! The opposite is true. We have real time data. The Nikkei goes up on talk of printing money, down on talk of not printing money, and goes *wild* on actual unexpected money printing. It's almost as if the market thinks printing money is awesome and has a rational expectations model. The bond market? The  rising interest rates? Not a peep.

Printing money wouldn't be prestigious! It would hurt bank independence! The opposite is true. Not printing money forced Prime Minister Shinzo Abe to threaten them into printing more money. They were seen as failures. Everyone respects the Bank of Australia because they *did* print more money.

This same vague fear, combined with trivial inconveniences, is what stops the other solutions, too.

Not only are these trivial fears that shouldn't stop us, *they're not even things that would happen.* When you try the thing, *almost nothing bad of this sort ever happens at all.*

At all. This is low risks of shockingly mild social disapproval. Ignore.

**These worries aren't real. They're in your head.**

**They're in my head, too. The voice of Pat Modesto is in your head. It is insidious. It says whatever it has to. It lies. It cheats. It is the opposite of useful.**

If someone else has these concerns, the concerns are in *their* head, whispering in *their* ear. Don't hold it against them. Help them.

Some such worries are real. They can point to real costs and benefits. Check! But they're mostly trying to halt thinking about the object level, to keep you from being the nail that sticks up and gets hammered down. When someone else raises them, mostly they're the hammer. The fears are mirages we've been trained and built to see.

You don't have that problem, you say? Great! Other people do have that problem. Sympathize and try to help. Otherwise, keep doing what you're doing, only more so. And congratulations.

VI

My practical suggestion is that if you do, buy or use a thing, and it seems like that was a reasonable thing to do, you should ask yourself:

Can I do *more* of this? Can I do this *better?* Put in more effort, more time and/or more money? Might that do the job better? Could that be a good idea? Could that be worth it? How much more? How much better?

Make a quick object level model of what would happen. See what it looks like. Discount your chances *a little* if no one does it, but only a little. Maybe half, tops. Less if those who succeeded wouldn't say anything. In some cases, the thing you're about to try *is actually done all the time, but no one talks about it.* If you suspect that, *definitely* try it.

You'll hear the voice. This isn't done. There must be a reason. When you hear that, get excited. You might be on to something.

If you're getting odds to try, try. Use the try harder, Luke! You can do this. Pull out More Dakka.

It's also worth looking back on things you've done in the past and asking the same question.

I've linked several times to the Challenging the Difficult sequence, but none of this need be difficult. Often all that's needed, but never comes, is an ordinary effort.

The bigger picture point is also important. These are *the most obvious things.* Those bad reasons stop *actual everyone* from trying things that cost little, on any level, with little risk, on any level, and that carry huge benefits. For other things, they stop *almost everyone.* When someone does try them and reports back that it worked, they're ignored.

Something possibly being *slightly* socially awkward, or causing a likely nominal failure, acts as a veto. Rationalizations for this are created as needed.

Adding that to the economic model of inadequate equilibria, and the fact that almost no one got as far as considering this idea at all, is it any wonder that you can beat 'consensus' by thinking of and trying object-level things?

Why *wouldn't* that work?

# In the presence of disinformation, collective epistemology requires local modeling

In Inadequacy and Modesty, Eliezer describes modest epistemology:

> How likely is it that an entire country—one of the world's most advanced countries—would forego trillions of dollars of real economic growth because their monetary controllers—not politicians, but appointees from the professional elite—were doing something so wrong that even a non-professional could tell? How likely is it that a non-professional could not just suspect that the Bank of Japan was doing something badly wrong, but be confident in that assessment?
>
> Surely it would be more realistic to search for possible reasons why the Bank of Japan might not be as stupid as it seemed, as stupid as some econbloggers were claiming. Possibly Japan's aging population made growth impossible. Possibly Japan's massive outstanding government debt made even the slightest inflation too dangerous. Possibly we just aren't thinking of the complicated reasoning going into the Bank of Japan's decision.
>
> Surely some humility is appropriate when criticizing the elite decision-makers governing the Bank of Japan. What if it's you, and not the professional economists making these decisions, who have failed to grasp the relevant economic considerations?
>
> I'll refer to this genre of arguments as "modest epistemology."

I see modest epistemology as attempting to defer to a canonical perspective: a way of making judgments that is a Schelling point for coordination. In this case, the Bank of Japan has more claim to canonicity than Eliezer does regarding claims about Japan's economy. I think deferring to a canonical perspective is key to how modest epistemology functions and why people find it appealing.

In social groups such as effective altruism, canonicity is useful when it allows for better coordination. If everyone can agree that charity X is the best charity, then it is possible to punish those who do not donate to charity X. This is similar to law: if a legal court makes a judgment that is not overturned, that judgment must be obeyed by anyone who does not want to be punished. Similarly, in discourse, it is often useful to punish crackpots by requiring deference to a canonical scientific judgment.

It is natural that deferring to a canonical perspective would be psychologically appealing, since it offers a low likelihood of being punished for deviating while allowing deviants to be punished, creating a sense of unity and certainty.

An obstacle to canonical perspectives is that epistemology requires using local information. Suppose I saw Bob steal my wallet. I have information about whether he actually stole my wallet (namely, my observation of the theft) that no one else has. If I tell others that Bob stole my wallet, they might or might not believe me depending on how much they trust me, as there is some chance I am lying to them. Constructing a more canonical perspective (e.g. a in a court of law) requires integrating this local

information: for example, I might tell the judge that Bob stole my wallet, and my friends might vouch for my character.

If humanity formed a collective superintelligence that integrated local information into a canonical perspective at the speed of light using sensible rules (e.g. something similar to Bayesianism), then there would be little need to exploit local information except to transmit it to this collective superintelligence. Obviously, this hasn't happened yet. Collective superintelligences made of humans must transmit information at the speed of human communication rather than the speed of light.

In addition to limits on communication speed, collective superintelligences made of humans have another difficulty: they must prevent and detect disinformation. People on the internet sometimes lie, as do people off the internet. Self-deception is effectively another form of deception, and is extremely common as explained in The Elephant in the Brain.

Mostly because of this, current collective superintelligences leave much to be desired. As Jordan Greenhall writes in this post:

> Take a look at Syria. What exactly is happening? With just a little bit of looking, I've found at least six radically different and plausible narratives:
>
> • Assad used poison gas on his people and the United States bombed his airbase in a measured response.
>
> • Assad attacked a rebel base that was unexpectedly storing poison gas and Trump bombed his airbase for political reasons.
>
> • The Deep State in the United States is responsible for a "false flag" use of poison gas in order to undermine the Trump Insurgency.
>
> • The Russians are responsible for a "false flag" use of poison gas in order to undermine the Deep State.
>
> • Putin and Trump collaborated on a "false flag" in order to distract from "Russiagate."
>
> • Someone else (China? Israel? Iran?) is responsible for a "false flag" for purposes unknown.
>
> And, just to make sure we really grasp the level of non-sense:
>
> • There was no poison gas attack, the "white helmets" are fake news for purposes unknown and everyone who is in a position to know is spinning their own version of events for their own purposes.
>
> Think this last one is implausible? Are you sure? Are you sure you know the current limits of the war on sensemaking? Of sock puppets and cognitive hacking and weaponized memetics?
>
> All I am certain of about Syria is that I really have no fucking idea what is going on. And that this state of affairs—this increasingly generalized condition of complete disorientation—is untenable.

We are in a collective condition of fog of war. Acting effectively under fog of war requires exploiting local information before it has been integrated into a canonical perspective. In military contexts, units must make decisions before contacting a central base using information and models only available to them. Syrians must decide whether to flee based on their own observations, observations of those they trust, and trustworthy local media. Americans making voting decisions based on Syria must decide which media sources they trust most, or actually visit Syria to gain additional info.

While I have mostly discussed differences in information between people, there are also differences in reasoning ability and willingness to use reason. Most people most of the time aren't even modeling things for themselves, but are instead parroting socially acceptable opinions. The products of reasoning could perhaps be considered as a form of logical information and treated similar to other information.

In the past, I have found modest epistemology aesthetically appealing on the basis that sufficient coordination would lead to a single canonical perspective that you can increase your average accuracy by deferring to (as explained in this post). Since then, aesthetic intuitions have led me to instead think of the problem of collective epistemology as one of decentralized coordination: how can good-faith actors reason and act well as a collective superintelligence in conditions of fog of war, where deception is prevalent and creation of common knowledge is difficult? I find this framing of collective epistemology more beautiful than the idea of a immediately deferring to a canonical perspective, and it is a better fit for the real world.

I haven't completely thought through the implications of this framing (that would be impossible), but so far my thinking has suggested a number of heuristics for group epistemology:

- Think for yourself. When your information sources are not already doing a good job of informing you, gathering your own information and forming your own models can improve your accuracy and tell you which information sources are most trustworthy. Outperforming experts often doesn't require complex models or extraordinary insight; see this review of Superforecasting for a description of some of what good amateur forecasters do.
- Share the products of your thinking. Where possible, share not only opinions but also the information or model that caused you to form the opinion. This allows others to verify and build on your information and models rather than just memorizing "X person believes Y", resulting in more information transfer. For example, fact posts will generally be better for collective epistemology than a similar post with fewer facts; they will let readers form their own models based on the info and have higher confidence in these models.
- Fact-check information people share by cross-checking it against other sources of information and models. The more this shared information is fact-checked, the more reliably true it will be. (When someone is wrong on the internet, this is actually a problem worth fixing).
- Try to make information and models common knowledge among a group when possible, so they can be integrated into a canonical perspective. This allows the group to build on this, rather than having to re-derive or re-state it repeatedly. Contributing to a written canon that some group of people is expected to have read is a great way to do this.
- When contributing to a canon, seek strong and clear evidence where possible. This can result in a question being definitively settled, which is great for the group's ability to reliably get the right answer to the question, rather than

having a range of "acceptable" answers that will be chosen from based on factors other than accuracy.

- When taking actions (e.g. making bets), use local information available only to you or a small number of others, not only canonical information. For example, when picking organizations to support, use information you have about these organizations (e.g. information about the competence of people working at this charity) even if not everyone else has this info. (For a more obvious example to illustrate the principle: if I saw Bob steal my wallet, then it's in my interest to guard my possessions more closely around Bob than I otherwise would, even if I can't convince everyone that Bob stole my wallet).

# Against Love Languages

The other day, a friend on facebook shared a post on love languages and asked their friends what their's were. I said that this did not fit my ontology for affection in a deep romantic relationship, and when someone asked me what ontoloy I used, I gave this short response (copied here so I can link people to it in the future).

Background: the notion of love languages is that there's five main ways humans express affection, and they are

- gift giving,
- quality time
- words of affirmation
- acts of service (devotion)
- and physical touch

The reason this is useful to think about (according to the wikipedia summary of the book) is that

> [P]eople tend to naturally give love in the way that they prefer to receive love, and better communication between couples can be accomplished when one can demonstrate caring to the other person in the love language the recipient understands. An example would be if a husband's love language is acts of service, he may be confused when he does the laundry for his wife and she doesn't perceive that as an act of love, viewing it as simply performing household duties, because the love language she comprehends is words of affirmation (verbal affirmation that he loves her). She may try to use what she values, words of affirmation, to express her love to him, which he would not value as much as she does. If she understands his love language and mows the lawn for him, he perceives it in his love language as an act of expressing her love for him; likewise, if he tells her he loves her, she values that as an act of love.

My comment is below.

---

It often seems to me like the seemingly important things people say in relationships, even good relationships, are the sorts of things you could say in any relationship. "It was really great to see you" "Let's do this again sometime" "Tell me about your day" "I love you".

Alternatively, the compliments I most enjoy giving and receiving, are the ones that could only be said to that person. To pull an example from my recent life, I'd moved into a place for ~3 months, and was leaving to return to university. I gave someone the following goodbye (I'll pretend their name is John):

"Hey John, when I found out I'd be living with you for 3 months, I was initially disappointed. I remembered you from meeting you during my CFAR workshop as someone who had a chipper act all the time, and had a faux enthusiasm for trying out new ideas for better epistemology and productivity. However, for the past 3 months, you've been that person day-in-day-out, and I realise it's not an act at all - you're genuinely enthusiastic about new ideas and are a really fun guy to be around. I'm looking forward to coming back and seeing you in a couple of months."

I think they said it was one of the favourite goodbye's they'd had.

At other times in my life, I've seen people disagree for hours before coming to understand each other. Getting an accurate model of how someone thinks and really feeling what it's like to be them from the inside, is a difficult and time-intensive task, but one of the things I want most in a close relationship is to be understood.

This summer just gone, I was hanging out with a different person, and they said to me "Huh, I've just noticed Ben that I feel safe and happy talking to you right now, yet you've totally disagreed with everything I've said. That's really cool, I don't find this experience with many other people." This made me feel very seen, because on gut, S1 level it's definitely something I optimise for, but I'd never really put it into words before, and nobody else had noticed (or at least told me) either. I felt a *strong* affection for the friend.

The reason the ontology feels wrong is that the example I gave above was, I suppose, 'words of affirmation', but if someone gave me a wordless gift/act of service, I might feel it too, if I felt understood and seen in a way that rarely happens. Relatively, the other things don't matter too much to me. Being seen and understood is my love language, and are the things that cause the most natural affection for me.

# Show LW: Diaspora Project Map and Preregistration Database

Hi everyone,

I and a handful of other folks have been compiling a list of different LessWrong Diaspora projects. I started it after someone asked me how they could volunteer for LessWrong and I didn't have a good answer for them.

It's optimized for people who are researching what sort of things have already been done in the community, or are looking for a place where they can help. The basic criteria for eligibility are:

- Being something I'm sure is a 'real' Diaspora project. i.e, started by a credible person who is a member of the Diaspora. (Membership is of course a little fuzzy, err on the side of inclusion.)
- Either a web presence or contact information.

Suggestions for additions should go in the comment box on that page. (I'll also accept them in this thread, but understand I'll probably only check back on it for the first three days or so.)

We've also put together a preregistration database. A preregistration database solicits plans for projects and experiments before their results are known. It helps with the problem where no one wants to talk about something that didn't work. Ideally it becomes part of a framework for helping people put together solid projects.

Right now that framework includes a Project Litmus Test that's meant to help people who want to do something not fail from the outset through poor planning. The current question set is based on what we think are common failure modes for Diaspora projects, but over time I'd like to refactor it based on empirical interaction with peoples real ideas. In the best case scenario a workshop would be developed that helps people go from concept to execution.

My hope is that such a framework could put a real dent in the chronic lack of coordination that you've already been told about a thousand times.

The Map

The Preregistation Database

# Cash transfers are not necessarily wealth transfers

Here's a common argument:

The problem with the poor is that they haven't got enough money. There's ample empirical evidence backing this up. Therefore, the obviously-correct poverty intervention is to simply give the poor cash. You might be able to do better than this, but it's a solid baseline and you should often expect to find that interventions are worse than cash.

There are technical reasons to be skeptical of cash transfers - which is why it is so important that the cash transfer charity GiveDirectly is carefully researching what actually happens when they give people cash - but until fairly recently, these objections seemed to me like abstruse nitpicks about an intervention that was almost analytically certain to be strongly beneficial.

But they're not just nitpicks. Cash isn't actually the same thing as human well-being, and the assumption that it can be simply exchanged into pretty much anything else is not obviously true.

Of course, saying "X is possibly wrong" isn't very helpful unless we have a sense of how it's likely to be wrong, under what circumstances. It's no good to treat cash transfers just the same as before, but be more gloomy about it.

I'm going to try to communicate an underlying model that *generates* the appropriate kind of skepticism about interventions like cash transfers, in a way that's intuitive and not narrowly technical. I'll begin with a parable, and then talk about how it relates to real-world cases.

# Two cities, two stadiums

In a world where the only thing people enjoy is baseball, there is a wealthy city. In this city, there is an excellently designed baseball stadium. The seats are amply sized and comfortable. Even the worst seats have a good view of the field. There are plentiful amenities, though the price of food and drink is high. There are awnings to block the rain. There are lights in case the game goes late.

Elsewhere, in a poor city, there is another baseball stadium. This one is shoddily built. The upper seats are tiny to cram in as many people as possible. The worst seats can hardly see the field, or are exposed to inclement weather. The aisles are too narrow. The mood is chaotic, and in the poorer areas of the stadium – but not in the expensive areas with the best views – people often get into fights. A smaller variety of cheaper, lower-quality food and drink is available. There is always a line for the bathroom, unless you paid for a box with a private one.

You happen to live in the rich city, and attend the rich stadium. You know two more things, from studying data within and between many cities in this world:

1. Within a city, the self-reported enjoyment of baseball scales logarithmically with individual wealth. In other words, if you ask people to rate their happiness on a scale from 1 to 10, then no matter how rich or poor someone is, they'll be the same distance on the scale from someone with twice their wealth. (For instance, if people with an annual income of $10,000 report an enjoyment level of 5.2 out of 10 on average, and people with an annual income of $20,000 report 5.5 out of 10, then people with an annual income of $40,000 report 5.8 out of 10.)
2. Between cities, self-reported enjoyment of baseball *also* scales logarithmically with (average) wealth.

Clearly, enjoyment is a simple positive function of money; the problem with the people in the poor stadium is that they haven't got enough of it. What's more, since the function is logarithmic, a dollar goes much farther for the poor than for the rich. So you send some to the poorest people in the poor stadium, hoping that it will do then some large multiple of the good it can do you.

Because you know that good-sounding charitable interventions often fail for surprising reasons, you decide to test your assumptions, by following up with the recipients of the cash transfers. What do you find?

It turns out people care about two things: the quality of their seat, and food.

Some recipients buy more, or better food. Because the prices are lower over there, the difference in quality is large for them, even though the money would only make a small difference to them. Other recipients buy their way into better seats. Again, since the seats in the poor stadium cost less than the seats in the rich stadium, they gain a lot more than you lose. Overall, it looks like everyone who receives money enjoys the game a lot more, so your belief in the merits of cash transfers has been confirmed.

Then you learn about a third statistical regularity that flies in the face of everything you've learned so far:

1. As individual cities get richer, average enjoyment of baseball games does not increase.

What's going on?

When someone pays for on a nicer seat, their experience of the game is improved. But if they've outbid someone else for that seat, that other person is now stuck in a worse seat. Buying someone a nicer seat in the same stadium does not improve the *average* game enjoyment, because – assuming fully booked stadiums – it makes someone else's experience worse, because they're now stuck in the bad seat. So it matters quite a lot how much of the variation in people's well-being comes from things that can easily be got more of, like food, and how much comes from locally scarce goods, like choice seats.

But richer cities have nicer stadiums! Doesn't that mean that if you transfer enough money to people in poor cities, the poor city stadiums will get better? Maybe not! Stadiums are pretty hard to change substantially once built. And maybe it was never the money that made the stadiums better; maybe cities that have their act together tend to be better both at making money, and at stadium-building. You don't have empirical reason to doubt this.

How should your strategy to improve people's baseball experiences take this into account?

First, cash transfers can still be of some value. For instance, the very poorest spectators may go hungry. If you send them money, and they use the money to buy food, they might enjoy the game a lot more without harming anyone else. But if you keep giving them money, eventually they'll have enough food. The only thing left to buy is a positional good: better seats.

Second, if there are high-wage cities with spare stadium seating, you might want to help people move there. They'll enjoy baseball games more, since even the comparatively bad seats will be better, without harming anyone else. Some might turn down the opportunity out of loyalty to their hometown team, but others might take you up on it.

Third, if there is a way you can improve your home stadium, this is an important good. If you're willing to learn foreign cultural norms and be genuinely curious about why poor people have worse stadiums, you might even be able to help them reform their institutions to get better stadiums built, which could make a big difference in the quality of their lives.

The second and third points go together well. If your town's stadium is at capacity, immigration doesn't help anyone enjoy a nicer game – but persuading the owner to *add more seats* can change that.

# How this applies to the real world

In The Price of Glee in China, Scott Alexander points out a few key facts about the happiness literature:

1. Within countries, self-reported well-being seems to scale logarithmically of average income. (Again, that means that each doubling of income corresponds to the same, constant increase in reported happiness points.)
2. Between countries, a similar relationship seems to hold.
3. Within countries, per capita GDP growth does not appear to lead to corresponding increases in well-being.

These are the same three statistical regularities I gave in the baseball hypothetical, and we should draw similar conclusions. GiveDirectly is an excellent cash-transfer charity. It's on the GiveWell top charities list in large part because it is taking such care to collect evidence about what actually happens when the global poor are simply given money.

To that end, I found this Vox article about GiveDirectly's basic income experiment interesting. In particular, it's interesting that the headline case is someone who sometimes went hungry, but is not going to spend the money on food. Instead, she's going to spend it on what's plausibly a positional good instead:

> She is expecting her third child very soon. […] I asked Jacklin if she’s ever gone the whole day without eating; she has. I asked when the last time this happened was. She told me, “Last week.”

But when the nonprofit GiveDirectly told her that it would give her, and every other adult in her village, a basic income payment of 2,280 Kenyan shillings (about $22) a month for the next 12 years, she knew immediately that she would not spend the money on food.

Her plan is to save the money and then use it to pay her children's school fees.

She is starving herself, while pregnant, in order to save for her kids' formal schooling. This looks like a really bad outcome.

# A brief digression on education

But education! She's investing in her kids! Isn't that good?

The education industry is already eating the developed world alive, with little apparent benefit. You might argue that there are diminishing returns – intuitively, education seems important. But, there's ample evidence that developing-world education often doesn't really improve people's productive capacity; it often seems no better than obedience school, when it's not being used purely as a certification of the ability to obtain a degree (and thus that you're in the right social class for certain positions).

For a particularly dramatic example, consider Scott's report on his experience in Haiti:

> Even if you're one of the lucky ones who can afford to go to school, your first problem is that the schools can't afford paper: one of our hosts told stories of Haitian high schoolers who were at the level of Western 5th graders because they kept forgetting everything: they couldn't afford the paper to take notes on!

The other problem is more systemic: schools teach everything by uninspired lecture even when it's completely inappropriate: a worker at our camp took a "computer skills" course where no one ever touched a computer: it was just a teacher standing in front of the class saying "And then you would click the word FILE on top of the screen, and then you'd scroll down to where it said SAVE, and then you'd type in a name for the file..." and so obviously people come out of the class with no clue how to use an actual computer. There's the money issue - they couldn't afford a computer for every student - and a cultural issue where actually going to school is considered nothing more than an annoying and ritualistic intermediate step between having enough money to go to school and getting a cushy job that requires education.

These are both entirely consistent with a story where education improves individual outcomes by inducting people into the "educated class". And Scott continues:

> We heard horror stories of people graduating from nursing school without even knowing how to take a blood pressure - a nurse who used to work at the clinic would just make her blood pressure readings up, and give completely nonsensical numbers like "2/19". [W]hen cornered this nurse absolutely insisted that the blood pressure had been 2/19 and made a big fuss out of it.

Likewise, bureaucrats also do not seem to be able to do the sorts of tasks we would expect formal schooling to qualify you for. One such task is alphabetization:

> Gail, our program director, explained that she has a lot of trouble with her Haitian office staff because they don't understand the concept of sorting numerically. Not just "they don't want to do it" or "it never occurred to them", but after months and months of attempted explanation they don't understand that sorting alphabetically or numerically is even a thing. Not only has this messed up her office work, but it makes dealing with the Haitian bureaucracy - harrowing at the best of times - positively unbearable.

Gail told the story of the time she asked a city office for some paperwork regarding Doctors Without Borders. The local official took out a drawer full of paperwork and looked through every single paper individually to see if it was the one she wanted. Then he started looking for the next drawer. After *five hours*, the official finally said that the paper wasn't in his office.

While investing in education to get a higher future salary is a net good in a simplified economic model, in practice it's often just buying what economists call *rents.* Even if those rents sometimes take the form of a job, it's clear from Scott's example that schooling is not enabling government bureaucrats to *create more value* for taxpayers than less-schooled people would be able to do – they're just outcompeting the unschooled for a fixed pool of jobs where they don't do much.

I suspect Haiti is especially bad for a bunch of reasons, but I don't know *how* exceptional it is. And it would be weird for marginal education to be bad in exceptional basket cases like Haiti, bad in well-to-do countries like the US, but good in the middle.

Here's another anecdote from a friend:

> I had the opportunity to observe one poor school in India and indeed, school literally didn't happen most days of the year for one reason or another. (Waiting for textbooks was one reason I remember them giving.) Also, the teachers were stealing the food the rotary club provided for free lunch. (Someone noticed and then they stopped, which I suppose is nice.)

And often when they showed up to school they were often grading papers from other schools for separate pay instead of teaching the kids.

At least they could get the teachers to show up to steal the lunches sometimes.

Banerjee and Duflo's *Poor Economics* is consistent with the view that this is a common problem in developing countries. In Chapter 4, they write:

> In 2002 and 2003, the World Absenteeism Survey, led by the World Bank, sent unannounced surveyors to a nationally representative sample of schools in six countries. Their basic conclusion was that teachers in Bangladesh, Ecuador, India, Indonesia, Peru, and Uganda miss one day of work out of five on average, and the ratio is even higher in India and Uganda. Moreover, the evidence from India suggests that even when teachers are in school and are supposed to be in class, they are often found drinking tea, reading the newspaper, or talking to a colleague. Overall, 50 percent of teachers in Indian public schools are not in front of a class at a time they should be. How are the children supposed to learn?
>
> In 2005, Pratham, an Indian NGO focused on education, decided to go one step further and find out what children were really learning. […] Close to 35 percent of children in the seven-to-fourteen age group could not read a simple paragraph (first-grade level) and almost 60 per-cent of children could not read a simple story (second-grade level). Only 30 percent could do second-grade mathematics (basic division). […]
>
> Unfortunately, India is not unique: Very similar results have been found in neighboring Pakistan, in distantKenya, and in several other countries. In Kenya, the Uwezo Survey, modeled on ASER, found that 27 percent of children in fifth grade could not read a simple paragraph in English, and 23 percent could not read in Kiswahili (the two languages of instruction in primary school). Thirty percent

> could not do basic division. In Pakistan, 80 percent of children in third grade could no tread a first-grade-level paragraph.

In addition, parents seem to view education as a way to buy a credential, as a ticket into an "educated class" job, not a way for their children to pick up valuable skills continuously:

> Parents seem to see education primarily as away for their children to acquire (considerable) wealth. The anticipated route to those riches is, for most parents, a government job (as a teacher, for example), or failing that, some kind of office job.

Parents thus would rather pay for the complete education of their highest-potential children, then pay for a mostly-complete education for all their children. (Poor Economics does claim that the all-or-nothing credentialist view is unrealistic and there are substantial gains for each year of education.)

I believe very strongly that *learning* is an important form of real capital, and it's entirely possible that formal schooling is more like this in the developing than in the developed world, but I haven't see the evidence, and at this point I think that if you try to justify the benefits of a social program by pointing to the bare fact that people are using it to buy more formal schooling, I think this is mostly adverse rather than positive evidence. (I'm more optimistic about educational programs like SOLE that try to route around the hierarchy, and focus exclusively on learning rather than credentials.)

# Some investment is real

While some investments people make to better their or their children's circumstances are positional, others seem much more like the sort of real investment we might hope for, under the usual economic framework. For instance, one common use of GiveDirectly's cash transfers is to buy a metal roof, which lasts much longer than the more commonly used and cheaper thatch roofs. The most conservative estimates GiveWell cites suggest that the annual return on investment for metal roofs is 7-14%, though the other estimates they report are substantially higher.

Importantly, the return on investment for a tin roof is not plausibly extractive. Even though purchasing the roof imposes a cost on the world (by increasing demand for the relevant inputs), it also reduces a cost (by reducing demand for thatch roof inputs), and the reduction seems to be substantially greater than the increase.

At least some recipients of cash transfers try to start businesses, which is maybe the paradigmatic case of an investment for which we should expect a return. But while locals know their local economy much better than I do, I am skeptical of a business plan where the basic income is being used to subsidize variable rather than fixed costs. As Vox reports:

> Samson [...] explained his plan to go into cage fishing at the lake. He'd already bought the fish and just needed to buy feed, and the feed — per a catalog he showed us — is expensive. So he's going to use the GiveDirectly money for that part of the operation.

Business acumen, like anything else, is a skill that can take time to develop - while there are serious disadvantages to making decisions for people from afar, we should

at least not expect that cash transfer recipients will *immediately* make reasonable business decisions.

# How should we think about cash transfers now?

I hope that most readers will be see that the point here is *not* that cash transfers are bad. The point is that when you look at the evidence about what happens when rich people send poor people money, good news won't look like a tedious confirmation of an obvious truth, but rather like an encouraging outcome where one should have been *actually uncertain* beforehand. And you should be able to recognize ambiguous or bad news as such, and not assume that it's good by definition.

I hope that it will also motivate some amount of additional justified skepticism of nominal rates of return, as estimates of the social value of an investment. You, as someone living in a rich country, might not think that your access to higher-wage jobs is a reliable measure of how much value you're able to provide to the world. (If you did think this, the most benevolent thing you could do would be to maximize your nominal wealth.) If you are in fact skeptical of the meaningfulness of your income as a metric, you should be similarly skeptical of the meaningfulness of variations in income of people in poor countries.

Finally, while abstractions like income levels and average self-reported happiness can be good starting points for generating hypotheses, I hope I've been somewhat persuasive that they are not adequate metrics for making philanthropic decisions - one has to engage with *what's concretely happening in the world.*

(Cross-posted on my personal blog and the Effective Altruism forum.)

# Mapping the Social Mind (Buttons)

I said before that normal people were like a vast wall of buttons. Each button triggers a response: out pops a slip of paper with a rehearsed mini-speech, or an idea about what to do. Very much like cached responses, each one being triggered as its corresponding concept is activated by a stimulus.

I say "minimum wage." That is the stimulus.

The brain of a normal person resolves the sensory data and activates a concept. The "minimum wage" concept in the brain lights up. That is the pushing of the button.

The person says "Oh, yes, did you hear about the effects observed in ____land? Just goes that show that minimum wage is good/bad, doesn't it!" That is the cached response, the slip of paper that comes out.

One of the mistakes I made before I understood this system is that I expected people to be less hypocritical. Better said (for they're not really being hypocrites, at least, not in the important sense of the word), I expected their expressed beliefs to give me clues to how they would behave.

For example, respectable people in American society will tell you that "family is the most important thing."

Ah, you might think (subconsciously). I shall hereafter expect their revealed preferences to favor family very highly. If their brother comes looking for some help, you might predict that help will gladly be given.

Hahaha, yeah, that's pretty silly, amiright?

Think of it this way. The brother asking for help is like one button being pushed, the brother-wants-help button. Mentioning "family values" pushes another button, the recite-respectable-mantra-about-family-values button. What happens when you push those buttons? The deeper question behind that is, What determines what gets written on the little slips of paper? What determines which thoughts get cached and which don't?

For normal humans, the answer is that their social experiences decide what will be on the papers. You can predictably expect to find written on their papers whatever response that they have, so far, found to give them the most prestige in their in-group. Responses that win the adoration and adulation of their audience are kept, and refinements are tested over time and made permanent if found favorable.

Now, it is clear, yes? Why would the slip of paper from one button tell you anything about the slip of paper of another button? They're completely unrelated, don't you see? The papers are not compared to each other; the one is not used to write the other, and so, the decision to help the brother or not is not determined by what the shiniest answer to "family values" questions is!

In reality, it's a bit more complicated. After all, you can use their answer to "family values" to infer their ingroup, then call to mind the list of that ingroup's values, and then use that to deduce how they migth behave when the brother asks for help. If the ingroup *both* says to recite line X in response to "family values" stimuli *and* says to

perform act Z in response to a brother, then you might indeed use the one answer to infer the other. But the *point* is that the connection between the two papers does NOT come from trying to build a consistent philosophy to live by. X doesn't actually tell you anything about Z! The causal connection is less direct, because the real drivers of all the acts and opinions, and of whether or not they end up connected or not, are social status and ingroup norms.

As such, the usefulness of their expressed "opinion" on the importance of family is no more useful in predicting their behavior towards their family and in no more direct way, than is observing their manner of dress. Both are equally connected to their behavior towards their family, which is to say, not that much. And the connection is that both indicate which ingroup they identify with, which is useful for predicting generabl behaviors. The connection is *not* because the one answer actually directly tells you something about the other on the mere and unimportant basis that the one is the abstract representation of the other.

In short, the normal, non-nerdy, social, political animal that is the human does not use abtract thoughts like nerds do; they are not used to tell you something that is true about all the members in the category that the abstract refers to (well, *doesn't* refer to, but seems to from a nerdy interpretation). Rather, they, as all other behaviors, are used to gain social status. It's perfectly consistent when seen from this angle, but also very inconsistent when judged from the angle of...let's see, how to say...

If you take their abstract statements, and you interpret them naively, literally...you interpret them *as if* they were meant to describe the qualities of a certain category of things...if you interpret them *that* way, then they say something which is not true at all, does not accurately describe the members of the category in the slightest.

But I feel dumb to have ever interpreted them that way! Couldn't I see that the illusion of contradiction was because I was taking a perfectly natural piece of social signaling and filtering it through a weird interpretive system that it was never meant for? It was really all my fault; I wasn't hearing what they were saying. They often said that I was misrepresenting them, but I never understood why, since I could repeat what they said almost verbatim and they'd agree that that was their stance.

Well, now I get it. My ears heard signals, but my mind heard abstract descriptions of reality. I thought in terms of describing reality, not garnering brownie points. It's weird to me, really, but so is quantum mechanics. This is how humans naturally behave, so I try not to get in the way of my own understanding by calling it weird.

# Rules of variety

At some point in high school I noticed an interesting thing about my choice of essay topics: it was definitely not allowed to be the same choice of topics that anyone else had.

One reason this seemed strange at the time was that I had never explicitly noticed this constraint or intended it, even though it was doing a lot of work. It was such a deeply assumed part of the basic rules of behavior that I didn't know it was there.

But it seemed extra strange once I noticed it, because it happened in the context of the rest of my classmates as one blithely ignoring this absolute law of reasonable behavior and all writing about the same hackneyed thing. Which probably wouldn't have even occurred to me to do if I had set out to write the most surprising essay I could. So, apparently other people didn't even have this rule, though it seemed so inbuilt in me.

And this wasn't just a failure to understand the rules of that assignment—I realized then that I had been assuming this constraint for every essay, and um, perhaps everything.

(In retrospect 'maybe you are trying to be different and other people aren't' looks like an obvious explanation for the perplexing fact that I was different and other people weren't. But I was used to knowing about things I was doing intentionally, and I was only trying to be different in the sense that I currently try not to murder people—it would be so wrong that it doesn't cross my mind as a possibility. But while unconscious, this is a very effective form of intention.)

Years later, I think other people actually do have this rule, or similar rules. They just vary by topic, and I happened to be unusual on the topic of high school essays. But I think implicit constraints like this are actually pretty common, and usually feel too natural to be noticed, even while they entirely warp our behavior. I don't mean 'assumptions that we don't notice' in general, but in particular ones about how similar or dissimilar out behavior should be to others. I rarely hear these things spoken of, except to remark when they are broken, without comment on what they actually are or consideration as to whether they should be there.

Some examples of actions I think you would avoid to at least some extent, or make an excuse for:

- Showing up in the same outfit as someone else
- Naming your children the same names as your friend's children
- Decorating your room exactly the same way as your housemate (a friend of mine actually moved to a different room in the same shared house, leaving his art behind for the appreciative incoming resident, and replacing it with identical art in his new room. This seems widely considered weird.)
- Using the same unusual adjective multiple times in the same article without it making an intentional point
- Answering 'how are you?' with the same contentful description of your state as the one you just heard, without comment
- Getting the same unusual car as your colleague
- Using a turn of phrase that has been used many times before
- Doing a thing that is trite, hackneyed, cliche, or stale

- Going on holiday to the same place your friend just did
- Doing a project that is basically the same as one someone else did, without it being connected to theirs
- Using the same stylistic touches that others use (e.g. even though xkcd is widely considered good, if I draw comics that look just like xkcd, it would be weird)
- Copying too many of anyone's personal habits when you are not trying to flirt weirdly with them
- Showing up to prom in the same car as someone else

This may all sound pretty unimportant. Ok, society has to support more dress variety than would otherwise be optimal. Worse things happen. But I suspect this also shows up in intellectual activities and strategic decisions. And having random unacknowledged rules driving decisions in those places strikes me as more terrifying.

For instance, discussing how surveyed machine learning researchers expect human-level AI further out now than they did before the recent ML boom, someone pointed out to me that of course people are going to be pessimistic now, because the interesting thinkers a couple of years ago were optimistic, so optimism is now boring. If that person is right that the opinions of a field on a topic as important as how imminently they are bringing about the end of human dominion are mostly determined by the dynamics of fashionable distances in opinion-space, I say we have a problem.

Other places I'd expect to see this:

- Aversion to working on too close a question to someone else in your vicinity, if you are not working with them
- Aversion to just straightforwardly agreeing with another intellectual rather than emphasizing differences
- Aversion to liking things that are too popular (contrarianism)
- Aversion to strategies that are too popular, even if that doesn't affect their effectiveness
- Not discussing topics once they are too commonly discussed, even if they are not resolved.

It's old news that opinions move according to fashion. So why is this interesting?

First, I think we usually think of this as a pressure for conformity—for a few thought leaders to choose ideas somewhat freely then all the thought sheeple to follow. I'm claiming there are also strong forces for variety. And these don't just cancel and give us freedom—they lead to a narrow band of appropriate choices. The next step in the dance has to be a certain distance from the last.

Secondly, since opinions following fashion has been pointed out in the past, it is weird to point it out again. But human memory and salience probably require it to be pointed out sometimes, if we are to actually remember it.

I'm not very confident about all this, beyond the more basic observations. But it leads me to an image of culture evolving like a fractal river delta, every piece curling off into several pieces that are the right distance from it and one another. Which is kind of how culture seems.

# TSR #7: Universal Principles

***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 #7: UniversaL Principles

SUMMARY

- Don’t forget about soft technologies.
- Examples of Places with principle oriented decision making
    - Bridgewater (Ray Dalio’s hedge fund)
        - Principles (Ray Dalio’s book on how he thinks about principles)
    - Book In a Box (unorthodox publishing company)
        - The Book In a Box Culture Doc
- Rules Concerning principles
    - Principles must state action.
    - Principles must have an antithesis.
    - Principles must overrule power.
- Hundreds of all the little decisions you have to make are automatically made by having strong principles.
- Guidance
    - Search for and define universal principles
    - Diagnose before prescription
    - Operationalize

Pulling it all together

1. Recognize that universal principles can be distilled, and that it’s worth doing.
2. Schedule enough time to explore, codify, create, test, and operationalize them.
3. Start by *diagnosing the problem* first, before jumping into solutions.
4. Codify the problem; write it down.
5. Explore different solutions until you find a set of solutions that are the food fit.
6. Codify the solutions; write those down.
7. Operationalize it—build tools and practices around ensuring the principle is put into action.
8. Entrain it—practice until it’s consistently happening in your life and your organization.

I notice that often I feel a tinsy bit confused when thinking about principles. I have accepted that having explicit principles and operating on them is something that has a good shot at making me more effective, yet I’ve always felt a bit hazy about what exactly principles are. Should they outline the sort of world states I'm striving for? Should they be strict policies that inform me on how to make decisions? Do I decide them or discover them?

I’d really recommend you read the BIAB (Book In a Box) Culture document that Sebastian uses examples from. Seeing a list of 10 well thought out, well described principles has done wonders to my understanding of how to usefully think about

principles. Seriously, read through it. Principles seem to be the sort of things that are best served by extensional definitions, rather than intensional ones.

Having read through BIAB's principles, see if you agree with these claims.

- Principles are decision making aids that attempt to occupy a certain sweet spot of specificity. Too specific and your principles become unwieldy and have to be updated too often. Too vague and your principles can't actually inform your actions.
- Principles act sort of like a lighthouse in the distance. They strongly filter out a lot possible directions to take towards your goals, yet they stop short of telling exactly how to get to where you're going.
- There are lots of little decisions to make whose answers become obvious in light of a good principle.
- A good principle is one that won't systematically lead you astray; the worst it will do is not give you enough guidance.

The main change in my thinking is this: I used to be bothered by the fact that being very rigourous and formal in defining a principle seemed to kill it a little bit. Now I'm okay with principles not being formaly and rigourously definied.

When BIAB says that one of their Principles is "Act Like an Owner", I now get why they don't need a long technical definition of what it *really means* to "Act Like an Owner". There is some cluster of ideas in idea-space, and what I now think about when I hear Act Like an Owner, is likely very similar to what Tucker and Zach are thinking, courtesy of them giving such excellent extensional examples. They've definitely given this some rigorous thought, but that doesn't mean that they need a rigorous definition.

Here are some questions you might find useful:

How many principles is "too many"? What are examples of good and bad rationalist principles? Do you have any principles that you've thought through and live by? Any one's that you haven't thought through yet still live by? If there's a principle in your mind that you haven't quite found the right words for, what are some good examples you could use to give it an extensional definition?

# Epistemic Spot Check: Full Catastrophe Living (Jon Kabat-Zinn)

*Full Catastrophe Living* is a little weird, because between the first edition and the second a lot of science came out testing the thesis. For this blog post, I'm reviewing the new, scienced-up edition of FCL. However I have ordered the older edition of the book (thanks, Patreon supporters and half.com) and have dreams of reviewing that separately, with an eye towards identifying what could have predicted the experimental outcome. E.g. if the experimental outcome is positive, was there something special about the model that we could recognize in other self-help books before rigorous science comes in?

I originally planned on fact checking two chapters, the scientific introduction and one of the explanatory chapters. Doing the intro was *exhausting* and demonstrated a consistent pattern of "basically correct, from a small sample size, finding exaggerated", so I skipped the second chapter of fact checking. I also skipped the latter two thirds of the book.

## Overview

You've probably heard about mindfulness, but just in case: mindfulness is a meditation practice that involves being present and not holding on to thoughts, originally created within Buddhism. Mindfulness Based Stress Reduction (MBSR) is a specific class created by the author of this book, Jon Kabat-Zinn. The class has since spread across the country; he cites 720 programs in the introduction. *Full Catastrophe Living* contains both a playbook for teaching the class to yourself, the science of why it works (I'm guessing this is new?), a section on stress, and followup information on how to integrate meditation into your life.

## Introduction

*Claim: Humans are happier when they focus on what they are doing than when they let their mind wander, which is 50% of the time.*

**Accurately cited, large effect size, possible confounding effects.** (PDF). The slope of the regression between mind wandering and mind not-wandering was 8.79 out of a 100 point scale, and the difference between unpleasant mind wandering and *any* mind not-wandering task was ~30 points. Pleasant mind wandering was exactly as pleasant as focusing on the task at hand. Focusing accounting for 17.7% of the between-person variation in happiness, compared to 3.2% from choice of task.

Some caveats:

- People's minds are more likely to wander when they're doing something unpleasant, and when they are having trouble coping with that unpleasantness. The study could be identifying a symptom rather than a cause.
- The study population was extremely unrepresentative, consisting of people who chose to download an iPhone app.

*Claim: Loss of telomeres is associated with stress and aging; meditation lengthens telomeres by reducing stress (location 404).*

**Research slightly more theoretical than is represented, but theoretical case is strong.** (Source). First, let's talk about telomeres. Telomeres are caps on the ends of all of your chromosomes. Because of the way DNA is copied, they will shorten a bit on every division. There's a special enzyme to re-lengthen them (telomerase), but leading thought right now is that stress inhibits it. Short telomeres are associated with the diseases of aging (heart issues, type two diabetes) *independent of chronological age.* This is hard to study because telomere length is a function of your entire life, not the last week, but is pretty established science at this point.

Mindfulness reduces stress, so it's not implausible that it could lengthen telomeres and thus reduce aging. The authors also present some evidence that negative mood reduces the activity of telomerase. This is a very strong theoretical case, but is not quite proven.

*Claim: Happiness research Dan Gilbert claims meditation is one of the keys to happiness, up there with sleep and exercise (location 461).*

**Confirmed** that Gilbert is a happiness researcher and said the quote cited, although I can't find where he personally researched this.

*Claim: "Researchers at Massachusetts General Hospital and Harvard University have shown, using fMRI brain scanning technology, that eight weeks of MBSR training leads to thickening of a number of different regions of the brain associated with learning and memory, emotion regulation, the sense of self, and perspective taking. They also found that the amygdala, a region deep in the brain that is responsible for appraising and reacting to perceived threats, was thinner after MBSR, and that the degree of thinning was related to the degree of improvement on a perceived stress scale." (location 502)*

**Accurate citation, but:** small sample size (16/26), and for the first study the effect size was quite small (1%) for regions of a priori interest, and the second had quite wide error bands (source 1) (source 2). However the book does refer to these findings as preliminary.

*Claim: "They also show that functions vital to our well-being and quality of life, such as perspective taking, attention regulation, learning and memory, emotion regulation, and threat appraisal, can be positively influenced by training in MBSR." (location 508).*

**Misleading**. These are really broad claims and no specific study is cited. However, source 2 above has the following quote: "The results suggest that participation in MBSR is associated with changes in gray matter concentration in brain regions involved in learning and memory processes, emotion regulation, self-referential processing, and perspective taking." This is a very carefully phrased statement indicating that mindfulness is in the right ballpark for affecting these things, but is not the same as demonstrating actual change.

*Claim: "Researchers at the University of Toronto, also using fMRI, found that people who had completed an MBSR program showed increases in neuronal activity in a brain network associated with embodied present-moment experience, and decreases in another brain network associated with the self as experienced across time. [...] This study also showed that MBSR could unlink these two forms of self-referencing, which usually function in tandem." (location 508).*

**Accurate citation, small sample size** (36) that they made particularly hard to find (source). I can't decipher the true size of the effect.

*Claim: Relative to another health class, MBSR participants had smaller blisters in response to a lab procedure, indicating lower inflammation (location 529).*

**True, but only because the other class *raised* inflammation** (source)**.** Also leaves out the fact that both groups had the same cortisol levels and self-reported stress. So this looks less like MBSR helped, and more like the control program was actively counterproductive.

For the record, this is where I got frustrated.

*Claim: "people who were meditating while receiving ultraviolet light therapy for their psoriasis healed at four times the rate of those receiving the light treatment by itself without meditating." (location 534)*

**Accurate citation (of his own work), small sample size** (pdf).

Claim: "we found that the electrical activity in certain areas of the brain known to be involved in the expression of emotions (within the prefrontal cerebral cortex) shifted in the MBSR participants in a direction (right-sided to left-sided) that suggested that the meditators were handling emotions such as anxiety and frustration more effectively. […]

This study also found that when the people in the study in both groups were given a flu vaccine at the end of the eight weeks of training, the MBSR group mounted a significantly stronger antibody response in their immune system"

**Accurate citation (of his own work), slightly misleading, small sample size.** Once again, he's strongly implying a behavioral effect when the only evidence is that MBSR touches an area of the brain. On the other hand, the original paper gets into *why* they make that assumption, so either it's correct or we just learned something cool about the brain.

*Claim: MBSR reduced loneliness and a particular inflammatory protein among the elderly (location 551).*

**Not statistically significant.** (source) More specifically; the loneliness finding was significant but uninteresting, since the treatment was "8 weeks with a regular social activity" and the control was "not." The inflammation finding had $p = .075$. There's nothing magic about $p < .05$ and I don't want to worship it, but it's not a strong result.

I also researched MBSR in general, and found it to have a surprisingly large effect on depression and anxiety.

# The Model

To the extent *Full Catastrophe Living* has a model, it's been integrated so fully into the cultural zeitgeist that I have a hard time articulating it. It could be summarized as "do these practices and some amount of good things from this list will happen to you." Which kills my hypothesis that having a good model is necessary to getting good results.

# You Might Like This Book If…

I don’t know. I found it a slog and only read the first third, but the empirical evidence is very much on mindfulness’s side and I don’t know what better thing to suggest.

Thanks to the internet for making it possible for me to do these kinds of investigations.

Thanks to Patreon supporters for giving me money.

# Philosophy of Numbers (part 2)

A post in a series of things I think would be fun to discuss on LW. Part one is here.

---

**I**

As it turns out, I asked my leading questions in precisely the reverse order I'd like to answer them in. I'll start with a simple picture of how we evaluate the truth of mathematical statements, then defend that this makes sense in terms of how we understand "truth," and only last mention existence.

Back to the comparison between "There exists a city larger than Paris" and "There exists a number greater than 17." When we evaluate the statement about Paris we check our map of the world, find that Paris doesn't seem extremely big, and maybe think of some larger cities.

We can use exactly the same thought process on the statement about 17: check our map, quickly recognize that 17 isn't very big, and maybe think of some bigger numbers or the stored principle that there is no largest integer. A large chunk of our issue now collapses into the question "Why does the map containing 17 seem so similar to the map containing Paris?"

<Digression>

> We use the metaphor of map and territory a lot, but let's take a moment to delve a little deeper. My "map" is really more like a huge collection of names, images, memories, scents, impressions, etcetera, all associated with each other in a big web. When I see the word "Paris" I can very quickly figure out how strongly that thing is associated with "city size," and by thinking about "city size" I can tell you some city names that seem more closely-associated with that than "Paris."
>
> "17" is a little trickier, because to explain how I can have associations with "17" in my big web of association, I also need to explain why I don't need a planet-sized brain to hold my impressions of all possible numbers you could have shown me.
>
> The answer is that there's not really a separate token in my head for "17," and not for "Paris" either. My brain doesn't keep a discrete label for everything, instead it stores and manipulates mental representations that are the collective pattern of lots of neurons, and therefore inhabit some high-dimensional space. For example, 17 and 18 might have mental representations that are close together in representation-space. And I can easily represent 87438 despite never having thought about that number before, because I can map the symbols to the right point in representation-space.

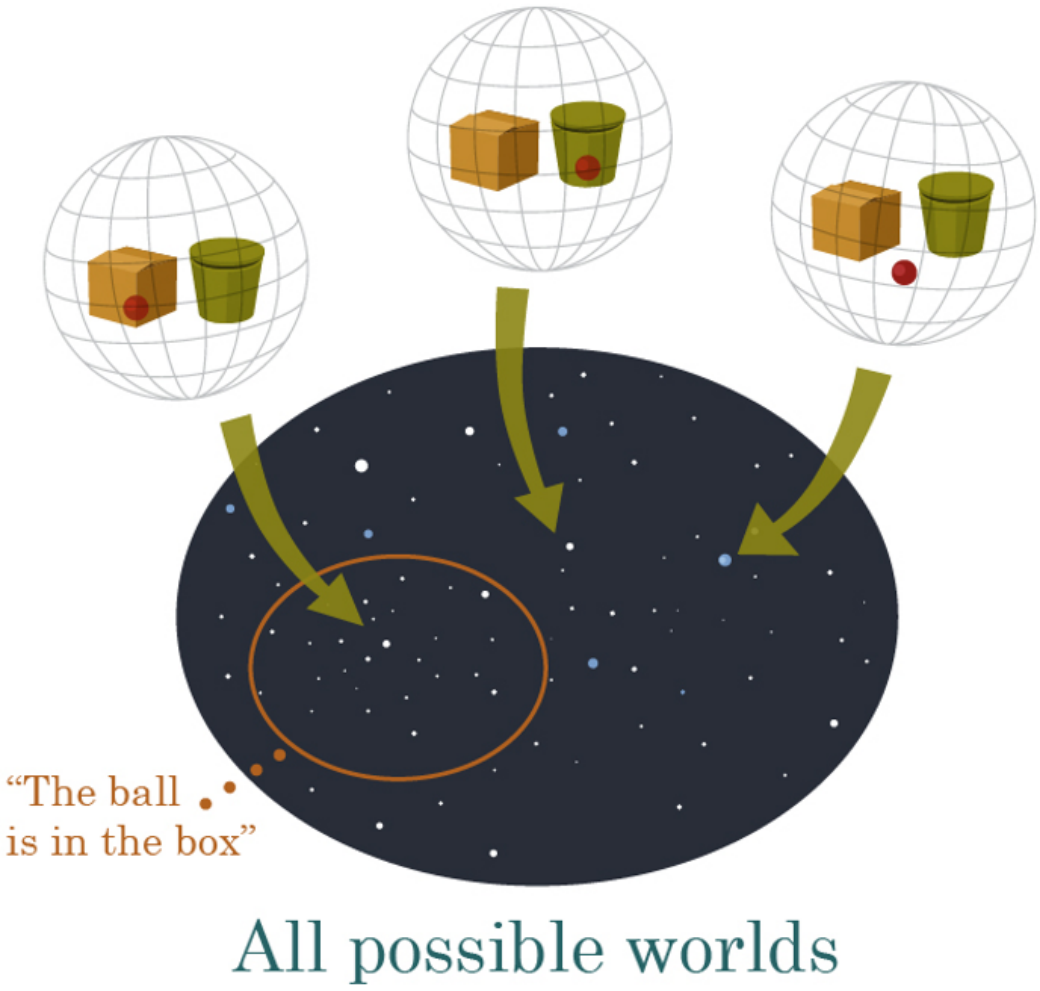
</Digression>

If we really do evaluate mathematical statements the same way we evaluate statements about our map of the external world, then that would explain why both evaluations seem to return the same type of "true" or "false." It's also convenient for evaluating the truth of mixed mathematical and empirical statements like "The number of pens on my table is less than 3 factorial." But we still need to fit this apparent-truth of mathematical statements with our conception of truth as a correspondence between map and territory.

**II**

An important fact about our models of the world is that they're capable of modeling things that aren't real. Suppose our world contains a red ball. We might hypothesize many different world-models and variations on models, each with a different past and future trajectory for

the red ball. Psychologically, this feels like we are imagining different possible worlds, at most one of which can be real.




To make a statement like "The ball is in the box" is to imply that we are in one specific fraction of the possible worlds. This statement is false in some possible worlds and true in others, but we should only endorse that the ball is in the box if, in our one true world, the ball is actually in the box.

Each statement about the red ball that we can evaluate as true or false can be thought of as defining a set of the possible worlds where that statement is true. "The volume of the ball contains a neutrino" is true in almost every world, while "The ball is in a volcano" is true in almost none. Knowing true statements gives us helps us narrow down which possible world we're actually in.

<Digression> More technically, knowing true statements helps us pick models that predict the world well. All this talk of possible worlds is a convenient metaphor. </Digression>

Moving closer to the point: "The ball has bounced a prime number of times" also defines a perfectly valid set of possible worlds. So. Does "3 is a prime number" define a set of possible worlds?

If we were really committed to answering "no" to this, we would have to undergo strange contortions, like being able to evaluate "The ball has bounced three times and the ball has bounced a prime number of times," but not "The ball has bounced three times and three is a prime number." Being able to compare the empirical with the abstract suggests the ability to compare the abstract with the abstract.

If we answer "yes," the set of possible worlds where 3 is a prime number seems like "all of them." (Or perhaps only almost all of them.) Math is then a bunch of tautologies.

But this raises an important problem: if mathematical truths are tautologous, then that would seem to render having a mental map of mathematics unnecessary - you can just evaluate statements purely on whether they obey their axioms. Conversely, if mathematical statements are always true or always false, then they're not useful, because learning them doesn't refine our predictions of the world. To resolve this apparent problem, we'll need a very powerful force: human ignorance.

Even though mathematical statements are theoretically evaluable from a small set of axioms, in practice that is much, much too hard for humans to do at runtime. Instead, we have to build up our knowledge of math slowly, associate important results with each other and with their real-world applications, and be able to place new knowledge in context of the old.

So it is precisely human badness at math that makes us keep a mental map of mathematics that's structured like our map of the world. The fact that our map doesn't start completely filled in also means that we can learn new things about math. It also leads directly into my last leading question from part one: why might we think numbers exist?

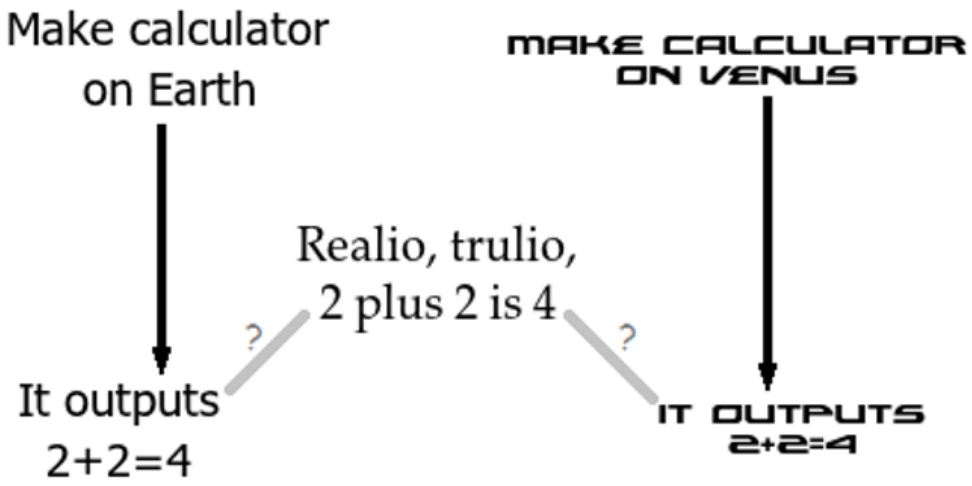
**III**

The reasons to feel like numbers exist are pretty similar to the reasons to feel like the physical world exists. For starters, our observations don't always turn out how we'd predict. The stuff that generates the predictions, we call belief, and the stuff that generates the observations, we call reality.

Sometimes, you have beliefs about mathematical statements even if you can't prove them. You might think, say, P!=NP, not by reasoning from the axioms, but by reasoning from the shape of your map. And when this heuristic reasoning fails, as it occasionally does, it feels like you're encountering an external reality, even if there's no causal thing that could be providing the feedback.

We also feel more like things exist when we model them as objective, rather than subjective. When we use our model of the world to imagine changing peoples' opinions about an objective thing, our model says that the objective thing doesn't change. Mathematical truths fulfil this property nicely - details left to the reader.

Lastly, things that we think exist have relationships with other elements in our map of the world. Things are associated with properties, like color and size - numbers definitely have properties. And although numbers are not connected to rocks in a causal model of the world, it seems like we say "2+2=4" because 2+2=4. But the "because" back there is not a causal relationship - rather it's an association our brain makes that's something like logical implication.




So maybe I do understand those mysterious links in LDT (artist's representation above) better than I did before. They're a toy-model representation of a connection that seems very natural in our brains, between different things that we have in the same map of the world.

---

**Epilogue**

I played a bit coy in this post - I talk a big game about understanding numbers, but here we are at the end and rather than telling you whether numbers really exist or not, I've just harped on what makes people *feel* like things exist.

To give away the game completely: I avoided the question because whether numbers "really exist" can end up getting stuck in the center node of the classic blegg/rube classifier. When faced with a red egg, the solution is usually not to figure out if it's "really a blegg or a rube." The solution is to be able to think about it as a red egg. And the even better solution is to understand the function of sorting these objects so that we can use categorizations in contexts where it's useful.

Understanding why we feel the way we do about numbers is really an exercise in looking at the surrounding nodes. The core claim of this article is that two things that normally agree - "should be a basic object in a parsimonous causal model of the world" and "can usefully be thought about using certain expectations and habits developed for physical objects" - diverge here, and so we should strive to replace tension about whether numbers "really exist" with understanding of how we think about numbers.

---

My aim was for a standard LW-ian view of numbers. I feel like I learned a lot writing this, and hopefully some of that feeling rubs off on the reader. (Thank you for reading, by the way.) I'll be back with something completely different next week.

# Comment on SSC's Review of Inadequate Equilibria

Copying over my comment from the SSC review, which otherwise may get lost in the fog of comments there.

---

Super fun review!

> *I found this part to be the biggest disappointment of this book. I don't think it grappled with the claim that the Outside View (and even Meta-Outside View) are often useful. It offered vague tips for how to decide when to use them, but I never felt any kind of enlightenment, or like there had been any work done to resolve the real issue here. It was basically a hit job on Outside Viewing.*

Conversely, I found the book gave short but excellent advice on how to resolve the interminable conflict between the inside and outside views – the only way you can: empiricism. Take each case by hand, make bets, and see how you come out. Did you bet that this education startup would fail because you believed the education market was adequate? And did you lose? Then you should update away from trusting the outside view here. Et cetera. This was the whole point of Chapter 4, giving examples of Eliezer getting closer to the truth with empiricism (including examples where he updated towards using the expert-trusting outside view, because he'd been wrong).

You quote "Eliezer's four pronged strategy" But I feel like his actual proposed methodology was in chapter 4:

> *Step one is to realize that here is a place to build an explicit domain theory—to want to understand the meta-principles of free energy, the principles of Moloch's toolbox and the converse principles that imply real efficiency, and build up a model of how they apply to various parts of the world.*
>
> *Step two is to adjust your mind's exploitability detectors until they're not always answering, "You couldn't possibly exploit this domain, foolish mortal," or, "Why trust those hedge-fund managers to price stocks correctly when they have such poor incentives?"*
>
> *And then you can move on to step three: the fine-tuning against reality.*

This is how you figure out if you're Jesus – test your models, and build up a track record of predictions.

You might respond "But telling me to bet more isn't an answer to the philosophical question about which to use" in which case I repeat: there isn't a way a priori to know whether to trust experts using the outside view, because you don't know how good experts are, and you need to build up domain-specific skills in predicting this.

You might respond "But this book didn't give me any specific tools for figuring out when to trust the experts over me" in which case I continue to be baffled and point you to the first book – Moloch's toolbox.

Finally, you might respond “Thank you Eliezer I’d already heard that a bet is a tax on bullsh*t, I didn’t require a whole new book to learn this” to which I respond that, firstly, I prefer the emphasis that “bets are a way to pay to find out where you’re wrong (and make money otherwise)” and secondly that the point of this book is that people are assuming way too quickly the adequacy of experts, so please make more bets in this particular domain. Which I think is a very good direction to push.

# Thinking as the Crow Flies: Part 2 - Basic Logic via Precommitments

# Preamble

In my last post I gave the philosophical prerequisites needed to understand my approach to logic and mathematics. Before getting into the subject at hand, here's a clarifying remark inspired by some comments on my last post. When I say that mathematics is a social activity, I mean it's an activity that is social. Politics is also a social activity. This means that saying, for example, that mathematics is incomplete or inconsistent makes as much sense as saying that politics is incomplete or inconsistent (in the same way a formal logic might be). It's very common for someone to make mention of mathematics having some property (e.g. Gödelian incompleteness) when, in fact, it's a specific logic (such as ZFC or the totality of formal logic) which has this property. This odd synecdoche causes people to frequently confuse a part of mathematics with the whole. No one logic or collection of logics constitutes all of mathematics. The intuitions used to justify an axiom are part of mathematics, but not part of any axiom system.

With that out of the way, the point of this post is to give explicit meanings to the most basic logical connectives.

# Logical Theories

A logical theory is a sort of collection of precommitments with two main kinds of judgments, $P\ \mathrm{Prop}$ (read "P is a proposition") and $P\ \mathrm{True}$ (read "P is true"). Within a logical theory, all but one of our precommitments will pertain to True. Note that any given precommitment will be about one particular thing, as previously explained. The propositional connectives we will initially deal with will be $\bot$ (falsum), $\top$ (verum), $\wedge$ (and), $\vee$ (or), and $\rightarrow$ (implies). Each will have a precommitment telling us precisely what's required for making a declaration of truth pertaining to that connective, and nowhere else will there be rules for making such declarations. Our one remaining precommitment will be the acting definition of Prop. Briefly, the precommitment for Prop says that $P$ is a proposition when we've said what's required for declaring $P\ \mathrm{True}$.

These precommitments will be variously referred to as "meaning explanations", descending from the usage of Martin-Löf, which explain the rules for reasoning about a judgment. For a judgment $J$, a meaning explanation should be a series of statements of the form:

> We may declare $J$ when ...

The meaning of the judgment $P\ \mathrm{Prop}$ is, then, as follows:

> We may declare $P\ \mathrm{Prop}$ when we understand the requirements for declaring $P\ \mathrm{True}$.

So, assuming that we follow our precommitments, we will only ever declare $P\ \mathrm{Prop}$ when we already have prescribed the rules by which we reason about $P$. This is part of the definition of $P$. Now, the judgment $P\ \mathrm{True}$ can only be declared in the case that we know the requirements for making such a declaration, meaning we should always be able to declare $P\ \mathrm{Prop}$ in such a circumstance.

The judgments True and Prop are called "categorical", signifying that they lack assumptions or generality. For the sake of addressing assumptions, we introduce a new form of judgment, which we call a "hypothetical judgment". A hypothetical judgment is made under a hypothesis. Our primary hypothetical judgment will be $\dfrac{J}{J'}$, pronounced "$J$ under the assumption $J'$". Its meaning explanation is as follows:

> We may declare $\dfrac{J}{J'}$ when we are allowed to declare $J$ after declaring $J'$.

Hypothetical judgment may be iterated, and $J_1\ |\ J_2$ will be used as notation for $\dfrac{J_1}{J_2}$

We introduce the notion of a "general judgment", which is judgment with respect to a variable, $|_x J$, pronounced "for an arbitrary x, $J$". The meaning explanation for general judgments is as follows:

> We may declare $|_x J$ when we are allowed to declare $J[x := E]$ (i.e. the substitution of $E$ for $x$ in the expression $J$) for any expression $E$.

We assume that the bar symbol binds as loosely as possible. Additionally, when making multiple general judgments, we'll use the notation $|_{x,y} J$ for $|_x |_y J$.

Note that this judgment relies essentially on the notion of substitution and variable binding. This, along with a full exposition on tokens and expressions will be addressed in my next post.

One thing worth clarifying is the difference between $\dfrac{P\ \mathrm{True}}{P\ \mathrm{True}}$ and $|_P \dfrac{P\ \mathrm{True}}{P\ \mathrm{True}}$. That first is more like a scheme for a judgment. It's not meant to be declared on its own. Rather, $P$ has to be replaced with a concrete expression by the person asserting the judgment. The second is a complete judgment which can be made as-is.

One principle worth noting is the following: in order to have declared $P\ \mathrm{True}$, we must have already declared the conditions under which such a declaration could be made. This means we must be able to declare $P\ \mathrm{Prop}$. This allows us to validate $\dfrac{P\ \mathrm{True}}{P\ \mathrm{Prop}}$.

I would like to digress at this point to critique the knowledge-theoretic interpretation of judgments. In *Type Theory and its Meaning Explanations*, for example, we read the

following;

> To know $P$ Prop is to know that $P$ is a proposition, which is to know what would count as a direct verification of $P$.

Though it may not be obvious before it's pointed out, this description is simply untrue. We may know what counts as a direct verification of $P$ without knowing that $P$ is a proposition.

There is an extra step missing, being the act of realization. In order to know $P$ Prop, we must first realize it. Let me demonstrate the importance of this distinction with an example. Here is a knowledge-theoretic meaning-explanation of sorts for prime factors;

> To know that m and n are prime factors of k is to know that m is prime, that n is prime, and that $m \times n = k$.

Based on that, let me ask you this. What are the prime factors of 91? If, by chance you know a few arithmetic tricks, then replace this with a number obtained from multiplying larger primes, but without looking it up, very few people would be able to answer this quickly. Now what's $7 \times 13$? Most people should be able to answer 91 fairly quickly, as well as recognize that 7 and 13 are prime numbers. Now, I'll ask again, what are the prime factors of 91?

Notice what knowledge was present. Knowledge that 7 and 13 were prime and that $7 \times 13 = 91$ was present, but knowledge that 7 and 13 are prime factors of 91 wasn't. In this case, the realization required computational effort. For more sophisticated mathematical notions, that computational effort may even be insurmountable.

It is not generally the case that knowing A and that A implies B allows us to know B. That would take extra effort in realization, and so the knowledge-theoretic account of meaning is flawed. I've replaced references to knowledge with an appeal to precommitments and allowances, and that seems to eliminate this problem while making the account more clear and closer to practice. Additionally, we side-step issues of computer knowledge. It will become clear that the system described in this work is implementable on a computer, but we would need to address the baggage of the knowledge-theoretic account. We are begged to answer for the assertion that a computer apparently knows something, bringing into question the meaning of knowledge itself. Thankfully, this problem evaporates along with the original references to knowledge.

# Basic Propositions

Now that the most commonly used judgments are out of the way, we can start populating our logical theories with specific examples of propositions.

Perhaps the thematically first proposition is trivial truth $\top$. It has a simple precommitment with only a single rule;

We may declare $\top$ True.

In other words, we can assert $\top$ True regardless of circumstance. We might say that the judgment of $\top$ True is trivial. Since we have a precommitment telling us the requirements for declaring $\top$ True, we are free to declare $\top$ Prop.

The simplest proposition that we may consider is falsum $\bot$. To define this as a proposition, we must specify the requirements for declaring $\bot$ True: there are none. In other words, we use the empty precommitment as part of our definition of $\bot$. Since we have a precommitment which, trivially, tells us the requirements for making these declarations, we are free to assert $\bot$ Prop.

Next, let's define conjunction. This will be the first time we rely on a hypothetical judgment in making a precommitment. Our goal is to declare the conditions under which we can declare P $\wedge$ Q True. Our full precommitment looks like;

We may declare P $\wedge$ Q True when we've declared P True and Q True.

Consider what's necessary in order to follow this precommitment, we will need to assume that we've previously stated the conditions for declaring P True and Q True, that is, we must already be allowed to declare P Prop and Q Prop. Upon realizing this, we verify the following judgment;

$$\frac{P\ \mathrm{Prop} \qquad Q\ \mathrm{Prop}}{P \wedge Q\ \mathrm{Prop}}$$

We can also declare the following, which summarizes the truth-theoretic content of our precommitment;

$$\frac{P\ \mathrm{True} \qquad Q\ \mathrm{True}}{P \wedge Q\ \mathrm{True}}$$

We can even characterize our precommitment as the smallest one validating this judgment.

I would also like to emphasize that the above precommitment is the entire definition of the proposition. We have introduced syntax and have specified the canonical methods for introduction. There is nothing more to be done.

It is more common in works of logic to "define" a piece of the theory in terms of some collection of inference rules which themselves are taken as a foundation which is not itself justified. By a meaning explanation account, such rules must be validated, are taken as derived, as something that must be realized, not given.

Throughout this work, I will make reference to a judgment being validated. A validation is not necessarily part of our logic proper. Rather, it is an account of how we may come to realize that a judgment can be made. We may employ rather sophisticated reasoning in making a realization, in validating a judgment.

Moving on to disjunction, we can state the precommitment for ∨ as;

We may declare P ∨ Q True when we've declared P True.

We may declare P ∨ Q True when we've declared Q True.

As with ∧, in order for this precommitment to be followed, we must have already declared what's required for declaring P True and Q True. That is, we must already be allowed to declare P Prop and Q Prop. Upon realizing this, we verify the following declaration;

$$\frac{P\ \text{Prop} \qquad Q\ \text{Prop}}{P \vee Q\ \text{Prop}}$$

The precommitment pertaining to ∨ can be characterized as the smallest precommitment validating the following two judgments;

$$\frac{P\ \text{True}}{P \vee Q\ \text{True}} \qquad \frac{Q\ \text{True}}{P \vee Q\ \text{True}}$$

To end this section, we should address implication and define the circumstances under which P → Q is true.

We may declare P → Q True when we've declared $\frac{P\ \text{True}}{Q\ \text{True}}$.

It's worthwhile to ponder what exactly is required to follow this precommitment, as it's more complicated than the previous ones. Consider the requirements for declaring $\frac{P\ \text{True}}{Q\ \text{True}}$. In order to state this, we must have stated two previous things. Firstly, we must have said what's required for declaring P True, that is, we must be able to declare P Prop. Secondly, under the assumption of P True, we must have stated the requirements for declaring Q True, that is, we must be able to declare Q Prop when we've already declared P True: we must be able to declare $\frac{P\ \text{True}}{Q\ \text{Prop}}$. From this reasoning, we can validate the following judgment;

$$\frac{P\ \text{Prop} \qquad \frac{P\ \text{True}}{Q\ \text{Prop}}}{P \to Q\ \text{Prop}}$$

Upon declaring the precommitment for →, we can make the following judgment, characterizing the truth-theoretic content of our precommitment;

$$\frac{\frac{P\ \text{True}}{Q\ \text{True}}}{P \to Q\ \text{True}}$$

# Witnesses

So far we have taken the Martin-Löf approach of prescribing the meaning of a proposition by specifying the conditions under which it can be introduced. This has been called the verificationist approach. In particular, it should be clear that, assuming we've accomplished the requirements for declaring $P\ \mathrm{Prop}$, in declaring $P\ \mathrm{Prop}$ we are saying that we have specified precisely what $P$ means.

One question that lingers is the notion of proof. I'll say variously that we can prove something, but I've not said what a proof actually is. That is, we've given meanings to the propositions themselves but not to their proofs. Fundamentally, the difference between logic and mathematics in general is that logic cares if something is true while mathematics cares why something is true. This may seem confusing at first, but we will see by the end of the fourth post of this series that the ability to do mathematics emerges from specifying and differentiating between proofs.

Before jumping in, we should discuss the notion of an elimination rule. To start, let's look at conjunction. There are two methods for eliminating $A \wedge B$. Firstly,

$$\frac{A \wedge B\ \ \mathrm{True}}{A\ \ \mathrm{True}}$$

How might one verify this? Consider the meaning of $A \wedge B\ \mathrm{True}$. In order to make that conclusion, it must have already been the case that we can conclude $A\ \mathrm{True}$. This is similarly the case with $B\ \mathrm{True}$, allowing us to verify the second elimination rule;

$$\frac{A \wedge B\ \ \mathrm{True}}{B\ \ \mathrm{True}}$$

There is a duel perspective to verificationism called pragmatism. In pragmatism, instead of prescribing the meaning of something in terms of how it's introduced, we prescribe meaning by how it's eliminated. So we'd have a precommitment for $\wedge$ which looks something like;

> We may declare $A\ \mathrm{True}$ when we may declare $A \wedge B\ \mathrm{True}$.
>
> We may declare $B\ \mathrm{True}$ when we may declare $A \wedge B\ \mathrm{True}$.

And it is precisely the elimination rules of $\wedge$ that contain the truth-theoretic content of the pragmatist precommitment.

Remember, a precommitment is only a list of rules for making judgments. We do intend to make cheap the ontological commitment to the ideas who's meaning is encoded in these precommitments, and to that end the verificationist approach seems, prima facie, more palatable. However, as we will see in a much later post on codata, the pragmatist approach is necessary to make sense of, for example, languages with infinite sentences. For now, we will focus on how these dual perspectives interact.

One thing that we can see by using both the introduction and elimination rules is that information is always traceable through the rules. Consider the derivation;

> Assume we have declared A True and B True. We can conclude $A \wedge B$ True. After that, we know that, in order to declare $A \wedge B$ True, we must have been able to declare A True, and so we do, we declare A True.

For any given declaration, there must be some justification for making it. Typically, such justifications boil down to appeals to some precommitment. It's obvious that the justifications used to justify both instances of A True are, on some level, the same. We can see the traceability of this justification through the derivation, that the only information present at the conclusion of the derivation was already present at the outset. This property is called *soundness*, and essentially acts as our justification for calling the idea of $\wedge$ coherent: we cannot come to know something which we did not already know purely by reasoning about $\wedge$, and so is the case with all logical and mathematical operators.

We can express this traceability of information with a reduction, of sorts, that the above derivation should reduce, in some sense, to

> Assume we have declared A True.

This, fundamentally, is the motivation for type theory. We can represent the rules of a commitment themselves with symbols in the same way we have done for the propositions. This also allows us to be precise with our claims of a validated derivation. We may say that "$M$ is a derivation validated by the precommitment for $P$", and be sure that this is true based on the form of $M$. We will denote this relation between $M$ and $P$ as $M : P$, pronounced "$M$ witnesses $P$". When such a judgment can be made, we call $P$ a type, and refer to our broader theory as a type theory rather than a logical theory.

We could try specifying a meaning like this;

> We may declare $M : P$ when $M$ is the name of a rule in the precommitment for $P$.

But, as stated in the introduction, this is insufficient. We may have, instead of a rule directly, a plan for deriving a rule or series of rules for witnessing $P$, an algorithm which, when run, will give a canonical witness in the same way $1 + 1$ is a natural number, but not a canonical one. Reduction to a canonical form will be represented by a judgment $M \downarrow M'$, pronounced "$M$ evaluates to $M'$".

> We may declare $M \downarrow M'$ when, in any declared judgment, $M$ may be replaced with another expression $M'$, creating a new judgment which can still be declared.

The meaning explanation for $P$ Prop must be accordingly modified to take into account the computational behavior of expressions. We may now have non-canonical propositions, that is, expressions that evaluate to a proposition but don't themselves have associated precommitments. We will call this judgment "$P$ prop". Furthermore, we need a notion of canonical truth. Within our precommitments defining canonical propositions, we will use the judgment $P$ True, and outside we will have $P$ true for non-canonical propositions. So we now have the three meaning explanations;

For Prop:

We may declare $P$ Prop when we understand what's required to declare $P$ True.

For true:

We may declare $P$ true when there is a $P'$ such that we may declare $P \downarrow P'$ and $P'$ True.

For prop:

We may declare $P$ prop when there is a $P'$ such that we may declare $P \downarrow P'$ and $P'$ Prop.

As an example we will update the justification for the following derivation:

$$\frac{P \ \text{prop} \qquad \dfrac{P \ \text{true}}{Q \ \text{prop}}}{P \rightarrow Q \ \text{prop}}$$

The verification of this rule must change to correspond to our new understanding of prop. $P \rightarrow Q$ is already in a canonical form. That is, we can declare $P \rightarrow Q$ prop so long as we can declare $P \rightarrow Q$ Prop. From there, we need to establish what we need in order to declare $\dfrac{P \ \text{true}}{Q \ \text{true}}$. In order to state this, we must have stated two previous things. Firstly, we must have said what's required for declaring $P'$ True, where we may declare that $P \downarrow P'$. That is, we must be able to declare $P$ prop. Secondly, under the assumption of $P$ true, we must have stated the requirements for declaring $Q'$ True, where we may declare that $Q \downarrow Q'$. That is, we must be able to declare Q prop when we've already declared $P$ true: we must be able to declare $\dfrac{P \ \text{true}}{Q \ \text{prop}}$.

We will also modify previous meanings to make them more flexable. For example;

We may declare $P \wedge Q$ True when we've declared $P$ true and Q true.

This allows our connectives to operate on non-canonical propositions.

We will make a new judgment, $P$ type, which modifies our prop judgment. The only real distinction is that types have names for each precommitment rule. We have two precommitments for the canonical and noncanonical forms.

For Type:

> We may declare $P$ Type when we understand the requirements for declaring $M ; P$ for all applicable expressions $M$.

For type:

> We may declare $P$ type when there is a $P'$ such that we may declare $P \downarrow P'$ and $P'$ Type.

When a name appears within a precommitment, we call it a *canonical* witness of the type. Here, ; is used to denote that something is a canonical witness. The : token, however, is intended to relate non-canonical expressions as well, so we have the further meaning explanation in the case that $M$ is non-canonical;

> We may declare $M : A$ when we have an $M'$ and $A'$ such that we may declare $M \downarrow M'$, $A \downarrow A'$, and $M' ; A'$.

We may now modify our meaning of $\wedge$ with a named witness;

> We may declare $(a, b) ; A \wedge B$ when we may declare $a : A$ and $b : B$.

Note that a and b do *not* have to be canonical. Even if a and b are non-canonical, $(a, b)$ *is* canonical, as it's subject to the rules of the precommitment.

It's useful to start declaring non-canonical witnesses at this point. We do this simply by declaring a sequence of tokens, along with reduction rules. For example we have;

$$\pi_1((a, b)) \downarrow a \qquad \pi_2((a, b)) \downarrow b$$

Note the new kind of meaning, signified by $\downarrow$. Note that these aren't the meanings themselves. Instead, we have a meaning like;

> In any declared judgment, $\pi_1((a, b))$ may be replaced with a, creating a new judgment which can still be declared.

This models the meaning of $\downarrow$, and allows us to validate $\pi_1((a, b)) \downarrow a$. However, that's quite long-winded, so I'll just state the reduction rules themselves.

Each reduction rule constitutes a new precommitment. We ought to be concerned by the ontological cost of such a commitment. We can, at least, see that so long as our source expression is not in any kind of canonical form, we're neither introducing a new idea nor

modifying an old one. If all of our reduction rules are of this form, the precommitment is free of cost, it doesn't make our ideas incoherent. We may think of the reduction rules as the precommitment defining the meaning of the operators they reduce. If we were to assert a reduction rule for an operator which already has a meaning explanation, such as $\wedge$, for instance, then we'd be breaking the precommitment which is its meaning explanation.

As long as they do not conflict with any previous precommitment, we may freely assert any reductions we want, including looping and exploding expressions. These declarations, however, may not interface well with our previous ones, they do have to answer to the meaning explanations of our theory in order to be useful. For example, we may want to validate a derivation such as this;

$$\frac{p : A \wedge B}{\pi_1(p) : A}$$

To validate this rule, we observe that, in order for $p : A \wedge B$ to have been concluded, it must reduce to a canonical $p' ; A \wedge B$ (by the meaning of : on non-canonical terms). The only canonical witnesses of $A \wedge B$ are of the form $(a, b)$, where $a : A$ and $b : B$. This brings our target judgment to;

$$\frac{a : A \quad b : B}{\pi_1((a, b)) : A}$$

Since we've declared that $\pi_1((a, b)) \downarrow a$, we may alter our goal to

$$\frac{a : A \quad b : B}{a : A}$$

which is easily declarable.

We use the expression $\pi_1((a, b))$ to stand for the proof tree expressed by the reasoning in the earlier derivation. The previous validation acts as a precise restatement of that derivation where our justifications were named by the witnesses. The reduction rule gives a precise description for how to extract the assumption we initially made.

# A First Pass at Lambda Expressions

In the case of implication, we can easily transform the previous meaning explanation for the sake of making $\rightarrow$ a type;

We may declare $\lambda x. y ; A \rightarrow B$ when we may declare $\frac{x : A}{y : B}$.

For example, we may validate the following judgment;

$$\lambda\, p\, .\, \pi_1\, (\, p\, )\ :\ (\, A\ \wedge\ B\, )\ \rightarrow\ A$$

To do so, we must be able to state

$$\frac{p : A \wedge B}{\pi_1(p) : A}$$

which we, in fact, have already done. However, there's an extra consideration which must be made. For an expression $\lambda x.\, y; P \to Q$ may be the case that y is a non-canonical witness of Q and that y actually mentions x. Consider that, in order for $\lambda x.\, y$ to truly be a proof of $A \to B$, we must be able to construct a witness of B out of x, **assuming only** that x is a witness of A. That is to say, if we know more about the specific structure of x, then it may not be fully general, it may be creating a witness of B out of a special case of A rather than As in general. As a consequence, the meaning explanation above fails to capture what we intend. To demonstrate this, consider the following meaning for Maybe(A);

We may declare Nothing; Maybe(A).

We may declare Just(a); Maybe(A) when we may declare $a : A$.

We'll consider these kinds of constructs more deeply later on, but for now, consider that a p such that $p : \text{Maybe}(A)$ could have been declared in two canonical ways; one which contains a witness to A, and one which doesn't. Now, consider the judgment;

$$\frac{\text{Just}(a) : \text{Maybe}(A)}{a : A}$$

Assuming our previous meaning explanation was correct, then it should be the case that $\lambda \text{Just}(a).\, a : \text{Maybe}(A) \to A$, but this obviously shouldn't hold in general. Lambda expressions should only be able to bind variables, treated as witnesses of unknown structure, but not expressions containing constants. As a consequence, the previously stated meaning explanation is insufficient for characterizing functionality, the type-theoretic generalization of implication will need something more powerful than hypothetical judgment. A simple solution presents itself. We may use a generalized judgment;

We may declare $\lambda x.\, y; A \to B$ when we've declared $\vert_x \frac{x : A}{y(x) : B}$.

where that y(x)signifies that x may be mentioned in y, but not in A or B. This kind of abstract binding will be dealt with in the next post, and functional contexts will be defined in the post after that.

In the case of propositional implication, we may validate the following elimination rule;

$$\frac{A \;\; \text{true} \qquad A \to B \;\; \text{true}}{B \;\; \text{true}}$$

Consider that, in order to have concluded $A \to B$ true, we must have been able to conclude $\frac{A \text{ true}}{B \text{ true}}$, rendering our goal;

$$\frac{A \text{ true} \qquad \frac{A \text{ true}}{B \text{ true}}}{B \text{ true}}$$

which is easily declarable.

Using this reasoning, we can make the following deduction;

> Assume we have declared A true and may declare $\frac{A \text{ true}}{B \text{ true}}$ by some means. We can, by the meaning of $\to$ conclude $A \to B$ true. From the elimination rule, we may conclude B true.

It's clear that $A \to B$ true is simply an internalization of the judgment $\frac{A \text{ true}}{B \text{ true}}$. We may conclude that the above deduction should, in some sense, be equivalent to;

> Assume we have declared A true and may declare $\frac{A \text{ true}}{B \text{ true}}$ by some means. By the meaning of hypothetical judgments, this means precisely that we may declare B true when we've declared A true, and so we do, we declare B true.

In the expression $\lambda x.\, y$, we have that y is a recipe of sorts for building our conclusion out of our assumption x. We define a new reduction rule, application which we will denote $ap(x; y)$ with the following reduction rule;

$$ap\ (\ \lambda\ x\ .\ \ y\ ;\ z\ )\ \ \downarrow\ \ y\ [\ x\ :=\ z\ ]$$

We may validate the following judgment;

$$\frac{x\ :\ A \qquad p\ :\ A\ \to\ B}{ap\ (\ p\ ;\ x\ )\ :\ B}$$

To do so, we note that we need to be able to state $ap(p; x) : B$ in the case that we've stated $x : A$ and $p : A \to B$, by the meaning of hypothetical judgments. In order to have stated $p : A \to B$, we must have some $p'$ such that $p'; A \to B$ and $p \downarrow p'$. The only canonical witness of $A \to B$ is of the form $\lambda a.\, b$, so our goal becomes $ap(\lambda a.\, b; x) : B$.

$$\frac{x\ :\ A \qquad \lambda\ a\ .\ \ b\ :\ A\ \to\ B}{ap\ (\ \lambda\ a\ .\ \ b\ ;\ x\ )\ :\ B}$$

We may reduce our goal, and also point out that, in order to have concluded $\lambda a.\, b; A \to B$, we must be able to declare $\left|_a \frac{a : A}{b : B}\right.$.

$$\frac{x : A \quad \left|_a \frac{a : A}{b : B}\right.}{b[a := x] : B}$$

In order to declare $\left|_a \frac{a : A}{b : B}\right.$, we must have been able to declare $\frac{a : A}{b : B}[a := E]$ for any expression E. In particular, this allows us to use x in place of E, allowing us to judge $b[a := x] : B$ This makes our goal

$$\frac{x : A \quad b[a := x] : B}{b[a := x] : B}$$

This is easily declarable.

Consider that, if we stuck with our previous meaning that didn't make use of hypothetical judgment, we would eventually get stuck on;

$$\frac{x : A \quad \frac{x : A}{b : B}}{b[a := x] : B}$$

We've solved that major problem with the meaning explanation, but we are begged to answer for our specification of variable usage. This can be clarified by an understanding of free variables, as described in the next post.

If we take $\lambda x.\, b$ and ap on their own, ignoring any idea that it has anything to do with logical formulas, we get a self-contained model for computation called the lambda calculus. It is, on its own, Turing complete. It was the second such system discovered which has this property all the way back in the 1930s by Alonzo Church. At that time, the connection to logic that I showed here was not known, only to be described in 1969 by William Alvin Howard. In many ways the lambda calculus can be thought of as the computational component of deduction, describing precisely, through substitution alone, how justifications in deductive inferences flow through an argument.

# Disjunction and Generalized Eliminators

We may straightforwardly alter the meaning precommitment for $\vee$ to incorporate witnesses;

We may declare $\mathrm{inl}(p); P \vee Q$ when we've declared $p : P$.

We may declare $\mathrm{inr}(q); P \vee Q$ true when we've declared $q : Q$.

Characterizing the elimination rule for ∨ is a bit more complicated. Consider the following judgment.

$$\frac{A \vee B \ \text{true} \qquad \overline{[A \ \text{true}]} \ C \ \text{true} \qquad \overline{[B \ \text{true}]} \ C \ \text{true}}{C \ \text{true}}$$

to validate it, we consider what requirements are necessary to conclude A ∨ B true. There are only two; either we declared A true or we declared B true. This splits our goal into two separate cases which we must validate individually;

$$\frac{A \ \text{true} \qquad \overline{[A \ \text{true}]} \ C \ \text{true} \qquad \overline{[B \ \text{true}]} \ C \ \text{true}}{C \ \text{true}} \qquad \frac{B \ \text{true} \qquad \overline{[A \ \text{true}]} \ C \ \text{true} \qquad \overline{[B \ \text{true}]} \ C \ \text{true}}{C \ \text{true}}$$

Both are easily declarable. Skipping some intuition, we declare the following;

$\text{case}\ (\ \text{inl}\ (\ x\ )\ ;\ a\ .\ p\ ;\ b\ .\ q\ )\ \downarrow\ p\ [\ a\ :=\ x\ ]$ $\quad \text{case}\ (\ \text{inr}\ (\ x\ )\ ;\ a\ .\ p\ ;\ b\ .\ q\ )\ \downarrow\ q\ [\ b\ :=\ x\ ]$

case acts as a non-canonical witness for elimination. To see why the computation rules hold, consider the following judgment;

$$\frac{d : A \vee B \qquad |_a \overline{[a : A]} \ p(a) : C \qquad |_b \overline{[b : B]} \ q(b) : C}{\text{case}\ (\ d\ ;\ a\ .\ p\ ;\ b\ .\ q\ )\ :\ C}$$

This should remind you of the introduction rule for implication. We first note that, in-order to declare d : A ∨ B we must have already a $d'$ ; A ∨ B where $d \downarrow d'$. To validate our goal, we do a similar case split on $d'$, which can only be of the form inl(x) where x : A or inr(y) where y : B. We'll focus on the first of these cases, since the other is more-or-less the same. We have our new goal;

$$\frac{x : A \qquad |_a \overline{[a : A]} \ p(a) : C \qquad |_b \overline{[b : B]} \ q(b) : C}{\text{case}\ (\ \text{inl}\ (\ x\ )\ ;\ a\ .\ p\ ;\ b\ .\ q\ )\ :\ C}$$

By the reduction rule for case, the conclusion reduces to p[a := x] : C. We can see from the assumption $|_a \overline{[a : A]} \ p(a) : C$ that we may conclude $p[a := x] : C$ which allows us to trivially finish validating our goal.

Note the . used in the syntax for case, similar to λx. y. This will be our standard notation for variable binding outside of contexts and generalized judgments. This will be discussed in detail in the next chapter.

We may alter our elimination implementation to rely on implication. We define a modified case;

$$\mathsf{casef}\ (\ f\ ;\ g\ ;\ \mathsf{inl}\ (\ x\ )\ )\ \downarrow\ \mathsf{ap}\ (\ f\ ;\ x\ ) \qquad \mathsf{casef}\ (\ f\ ;\ g\ ;\ \mathsf{inr}\ (\ x\ )\ )\ \downarrow\ \mathsf{ap}\ (\ g\ ;\ x\ )$$

We may make a similar judgment through the fact that we can internalize generalized judgments using →.

$$\frac{d\ :\ A\ \vee\ B \qquad f\ :\ A\ \to\ C \qquad g\ :\ B\ \to\ C}{\mathsf{casef}\ (\ f\ ;\ g\ ;\ d\ )\ :\ C}$$

We can employ a similar trick to unify our two rules for ∧ into a single generalized elimination rule. For ∨, we needed two functions to account for all possibilities. The issue with the ∧ eliminators is not that they don't account for all possibilities, after all, there is only one, the pair. The issue is that neither one on their own accounts for all data. If we look at $\pi_1$;

$$\frac{(\ a\ ,\ b\ )\ :\ A\ \wedge\ B}{\pi_1\ (\ (\ a\ ;\ b\ )\ )\ :\ A}$$

we notice that it throws away half of our pair. Likewise with $\pi_2$. We need to have functions which account for all possibilities and all data if we want to have a unified eliminator.

Generically, we need a function which potentially operates on all internal data of our type constructor. For ∧ we only have the two types it has as input. So, to eliminate a A ∧ B in general, we need a single function f : A → (B → C). To figure out what the return value will be, we just issue the components of the pairs to f. So we have;

$$\pi_{\ f}\ (\ f\ ;\ (\ a\ ,\ b\ )\ )\ \downarrow\ \mathsf{ap}\ (\ \mathsf{ap}\ (\ f\ ;\ a\ )\ ;\ b\ )$$

From which we can verify;

$$\frac{p\ :\ A\ \wedge\ B \qquad f\ :\ A\ \to\ (\ B\ \to\ C\ )}{\pi_{\ f}\ (\ f\ ;\ p\ )\ :\ C}$$

We can recover the original eliminators by setting f to be $\lambda x.\,\lambda y.\,x$ and $\lambda x.\,\lambda y.\,y$ since $\pi_f(\lambda x.\,\lambda y.\,x;(a,b)) \downarrow a$ and $\pi_f(\lambda x.\,\lambda y.\,y;(a,b)) \downarrow b$.

Furthermore, we can obtain $\pi_f$ from our old eliminators by defining it as $\lambda p.\,\lambda f.\,\mathrm{ap}(\mathrm{ap}(f;\pi_1(p));\pi_2(p))$. This is just a different form of an eliminator we already had, not a more powerful principle.

We could also make a generalized eliminator like so;

$$\pi\ (\ x\ .\ y\ .\ f\ ;\ (\ a\ ,\ b\ )\ )\ \downarrow\ f\ [\ x\ :=\ a\ ]\ [\ y\ :=\ b\ ]$$

It's this sort of eliminator which we'll make most often as we move onto mathematics in general.

# Verum and Falsum

It's far easier to deal with verum and falsum when it comes to witnesses. The meaning explanation for trivial truth is simply;

> We may declare tt; $\top$.

and the meaning explanation for trivial falsehood is still the empty precommitment.

For the elimination of $\top$, we consider the following;

$$\frac{\dfrac{\top \quad \text{true}}{A \quad \text{true}}}{A \quad \text{true}}$$

Consider what's necessary to declare $\dfrac{\top \ \text{true}}{A \ \text{true}}$. We would need to be able to declare A true whenever we can declare $\top$ true. By the meaning of $\top$, we may always declare $\top$ true, meaning that we must always be able to declare A true. We reduce our goal to;

$$\frac{A \quad \text{true}}{A \quad \text{true}}$$

which is trivially verifiable. Using this, we can construct the following;

$$\text{triv}\ (\ t\ .\ a\ )\ \downarrow\ a\ [\ t\ :=\ \text{tt}\ ]$$

Using a similar strategy to before, we may verify;

$$\frac{\left|_{t}\ \dfrac{t\ :\ \top}{a\ (\ t\ )}\ :\ A\right.}{a\ [\ t\ :=\ \text{tt}\ ]\ :\ A}$$

Of course, this is somewhat trivial, as the generalized judgment which we're assuming covers tt. All this leads us to conclude that we can't really eliminate out of $\top$ since it doesn't contain any information.

Regarding $\bot$, we have something more interesting. Consider the following;

$$\frac{\bot \quad \text{true}}{A \quad \text{true}}$$

how might this be verified? Consider all the cases from which we can conclude $\bot$ true. Well, there are none, we're done. We've verified this judgment for all the cases we may consider.

Some reader's unfamiliar with logic may find this reasoning a bit difficult to wrap they're head around, but it's called the principal of explosion, and is completely consistent with our previous forms of reasoning. For $\vee$ we needed to verify our elimination on two different cases. For $\top$, it was one case. For $\bot$, it's zero, allowing us to trivially finish verification.

For the lack of witnesses to $\bot$, we may verify, through similar reasoning as before, the following;

$$\frac{b : \bot}{b : A}$$

# Defined Connectives

There are a handful of connectives which I've neglected to give meaning explanations to. Most significantly are negation, $\neg$ and the biconditional $\leftrightarrow$. They are defined in terms of what we have already defined. We simply assert the following;

$$\neg A \downarrow A \to \bot \qquad A \leftrightarrow B \downarrow (A \to B) \wedge (B \to A)$$

Starting with the biconditional, we can validate the following judgments;

$$\frac{e : A \leftrightarrow B}{\pi_1(e) : A \to B} \qquad \frac{e : A \leftrightarrow B}{\pi_2(e) : B \to A} \qquad \frac{e : A \leftrightarrow B}{ap(\pi_1(e); a) : B} \qquad \frac{e : A \leftrightarrow B}{ap(\pi_2(e); b) : A}$$

which sums up most use cases for $\leftrightarrow$.

$\neg$ is more interesting. We can investigate it by considering how excluded middle interacts with the logic I've so far presented. Excluded middle states that every proposition is either true or false, that $A \vee \neg A$ true. This is NOT declarable in the logic I set out. Consider how this might interact with our proof theory. If it were declarable then there should be some uniform algorithm which, given any arbitrary proposition A, would return either inl(a) where $a : A$ or inr(na) where $na : \neg A$. There are, indeed, some logics which have this property, which will be discussed in a later post. For now, we can prove $\neg\neg(A \vee \neg A)$ true. How would we verify this judgment?

We can reduce $\neg\neg(A \vee \neg A)$ to $((A \vee (A \to \bot)) \to \bot) \to \bot$. To prove this, we start with a lambda term $\lambda x.?_1$, where $x : (A \vee (A \to \bot)) \to \bot$, and we need to find a $?_1 : \bot$. Well, if we constructed a proof of $A \vee (A \to \bot)$, then we could apply it to x to get our $?_1$. The only way we can get a proof of a $\vee$ is to use either inl or inr. Off the bat, there's no clear way to give a proof of A, but maybe there's a way to prove $A \to \bot$, so let's go with inr.

inr wants a proof of $A \to \bot$, so we start with a lambda term $\lambda y.\, ?_2$, where $y : A$ and we need to find a $?_2 : \bot$. Hey, now we have a proof of A, so we have $\mathsf{inl}(y) : A \vee (A \to \bot)$. We can issue that into x to get $\mathsf{ap}(x; \mathsf{inl}(y)) : \bot$, which is our $?_2$.

Rolling that into our lambda term, we have $\lambda y.\, \mathsf{ap}(x; \mathsf{inl}(y)) : A \to \bot$. We then have $\mathsf{inr}(\lambda y.\, \mathsf{ap}(x; \mathsf{inl}(y))) : A \vee (A \to \bot)$. We can issue this to x, obtaining $\mathsf{ap}(x; \mathsf{inr}(\lambda y.\, \mathsf{ap}(x; \mathsf{inl}(y)))) : \bot$, which is our $?_1$ goal. Rolling this into our outermost lambda term, we finally arrive at the following judgment;

$$\lambda\ x\ .\ \mathsf{ap}\ (\ x\ ;\ \mathsf{inr}\ (\ \lambda\ y\ .\ \mathsf{ap}\ (\ x\ ;\ \mathsf{inl}\ (\ y\ )\ )\ )\ )\ :\ \neg\ \neg\ (\ A\ \vee\ \neg\ A\ )$$

which contains all the information necessary to recover our reasoning, compressed into a single line. More importantly, that expression can itself be reasoned about. It's not just an elegant presentation of our justification, it's data as well.

# Empirical philosophy and inversions

A regular installment. Index is here.

---

This post is in large part a linkpost for an excellent talk on experimental philosophy, given by Ned Block (don't be put off by the title): https://www.youtube.com/watch?v=6lHHxcxurhQ . Apologies to those who dislike videos, but (especially at 1.5x or 2x speed) it's faster and more fun than reading a bunch of his papers, I swear.

Here's an example: one experiment Block talks about sticks electrodes to peoples' heads, and then subtly shows them a geometrical shape while they're doing another task. In the after-experiment report, some participants report they noticed the shape, and their electrode data can then be reviewed to see what brain activity was neccessary for noticing the shape even if they didn't know they had to notice it. It turns out, your brain doesn't need to be very active for you to be able to recall seeing the shape.

Block uses results like this to defend his thesis about the richness of conscious perception, and how early in the brain's perceptual systems activity can be experienced consciously. But this forms an interesting contrast with a deflationary view of consciousness.

Our agent of deflation is Marvin Minsky. Here's a video of him being deflationist. He has a favorite point, which is that people associate consciousness with lots of tasks, like being able to remember smells, or being able to imagine applying verbs to nouns, et cetera, but that this grouping is a human-made category, and thinking about these things as a group can get in the way of understanding them. The stuff we call conscious activity can, he says, be broken up into lots of sub-processes like smelling and abstract-verb-imagining that have a strong internal coherence but not much overlap with each other.

Which brings us back to Ned Block and consciousness of perception. It's possible to look at the several experiments Block talks about, not as different probes of a unified consciousness, but as probes of several functions of the brain that fall under the umbrella of consciousness.

Another of Block's examples is presenting different images to each eye, and using eye-tracking to determine which image the subject is experiencing. It's natural and effortless for us to think that this sort of consciousness is the same thing as the consciousness of remembering the shape, form the first example. It takes a weird, effortful inversion of perspective for me to think about what if would be like if the brain functions determining the two experiments had very little overlap.

This is the reason I linked to Dennett's article on the intentional stance earlier - Minsky's view can be thought of as deliniating a "conscious stance" as separate from a "process stance." In this view, consciousness is just a convenient way to predict mental things without looking too close. And so from this view, the kinds of brain activity the empirical philosophers are trying to pin down - where do you store representation of things you see, how fast do you put those representations into long-term memory, what where do you figure out which eye's signals are dominant in

determining your representation of the visual field, et cetera - are actually at a level of description below consciousness.

You may already be grumbling about the distinction between hard and easy problems of consciousness. These grumblings are fair, and we will get to that later. I just thought this was too much fun to not share.

# Oracle paper

Available on the arXiv, my paper on two types of Oracles (AIs constrained to answering questions only), and how to use them more safely.

> An Oracle is a design for potentially high power artificial intelligences (AIs), where the AI is made safe by restricting it to only answer questions. Unfortunately most designs cause the Oracle to be motivated to manipulate humans with the contents of their answers, and Oracles of potentially high intelligence might be very successful at this. Solving the problem, without compromising the accuracy of the answer, is tricky. This paper reduces the issue to a cryptographic-style problem of Alice ensuring that her Oracle answers her questions while not providing key information to an eavesdropping Eve. Two Oracle designs solve this problem, one counterfactual (the Oracle answers as if it expected its answer to never be read) and one on-policy (limited by the quantity of information it can transmit).

# Calling Bullshit – Lectures and Readings on Evaluating Scientific Research

This is a linkpost for http://callingbullshit.org/index.html

This is a relatively recent course from UW which has all of the lecture videos and reading materials now available online. It's very useful for learning how to evaluate scientific research without having enough knowledge to evaluate the actual content of the research papers.

# 2017 AI Safety Literature Review and Charity Comparison

*Summary: I review a significant amount of 2017 research related to AI Safety and offer some comments about where I am going to donate this year. Cross-posted from here upon request.*

## Contents



## Introduction

Like last year, I've attempted to review the research that has been produced by various organisations working on AI safety, to help potential donors gain a better understanding of the landscape. This is a similar role to that which GiveWell performs for global health charities, and somewhat similar to an securities analyst with regards to possible investments.  It appears that once again no-one else has attempted to do this, to my knowledge, so I've once again undertaken the task. While I've been able to work significantly more efficiently on this than last year, I have been unfortunately very busy with my day job, which has dramarically reduced the amount of time I've been able to dedicate.

My aim is basically to judge the output of each organisation in 2017 and compare it to their budget. This should give a sense for the organisations' average cost-effectiveness. Then we can consider factors that might increase or decrease the marginal cost-effectiveness going forward. We focus on organisations, not researchers.

Judging organisations on their historical output is naturally going to favour more mature organisations. A new startup, whose value all lies in the future, will be disadvantaged. However, I think that this is correct. The newer the organisation, the more funding should come from people with close knowledge. As organisations mature, and have more easily verifiable signals of quality, their funding sources can transition to larger pools of less expert money. This is how it works for startups turning into public companies and I think the same model applies here.

This judgement involves analysing a large number papers relating to Xrisk that were produced during 2017. Hopefully the year-to-year volatility of output is sufficiently low that this is a reasonable metric. I also attempted to include papers during December 2016, to take into account the fact that I'm missing the last month's worth of output from 2017, but I can't be sure I did this successfully.

This article focuses on AI risk work. If you think other causes are important too, your priorities might differ. This particularly affects GCRI and CSER, who both do a lot of work on other issues.

We focus virtually exclusively on papers, rather than outreach or other activities. This is party because they are much easier to measure; while there has been a large increase in interest in AI safety over the last year, it’s hard to work out who to credit for this, and partly because I think progress has to come by persuading AI researchers, which I think comes through technical outreach and publishing good work, not popular/political work.

My impression is that policy on technical subjects (as opposed to issues that attract strong views from the general population) is generally made by the government and civil servants in consultation with, and being lobbied by, outside experts and interests. Without expert (e.g. top ML researchers at Google, CMU & Baidu) consensus, no useful policy will be enacted. Pushing directly for policy seems if anything likely to hinder expert consensus. Attempts to directly influence the government to regulate AI research seem very adversarial, and risk being pattern-matched to ignorant opposition to GM foods or nuclear power. We don't want the 'us-vs-them' situation, that has occurred with climate change, to happen here. AI researchers who are dismissive of safety law, regarding it as an imposition and encumbrance to be endured or evaded, will probably be harder to convince of the need to voluntarily be extra-safe - especially as the regulations may actually be totally ineffective. The only case I can think of where scientists are relatively happy about punitive safety regulations, nuclear power, is one where many of those initially concerned were scientists themselves.  Given this, I actually think policy outreach to the general population is probably negative in expectation.

The good news on outreach this year is we haven’t had any truly terrible publicity that I can remember, though I urge organisations to remember that the personal activities of their employees, especially senior ones, reflect on the organisations themselves, so they should take care not to act/speak in ways that are offensive to those outside their bubble, and to avoid hiring crazy people.

Part of my motivation for writing this is to help more people become informed about the AI safety landscape so they can contribute better with both direct work and donations. With regard donations, at present Nick Beckstead, in his role as both Fund Manager of the Long-Term Future Fund and officer with the Open Philanthropy Project, is probably the most important financer of this work. He is also probably significantly more informed on the subject than me, but I think it's important that the vitality of the field doesn't depend on a single person, even if that person is awesome.

# The Machine Intelligence Research Institute (MIRI)

MIRI is the largest pure-play AI existential risk group. Based in Berkeley, it focuses on mathematics research that is unlikely to be produced by academics, trying to build the foundations for the development of safe AIs.

Their agent foundations work is basically trying to develop the correct way of thinking about agents and learning/decision making by spotting areas where our current models fail and seeking to improve them. Much of their work this year seems to involve trying to address self-reference in some way - how can we design, or even just model, agents that are smart enough to think about themselves? This work is technical, abstract, and requires a considerable belief in their long-term vision, as it is rarely locally applicable, so hard to independently judge the quality.

In 2016 they announced they were somewhat pivoting towards work that tied in closer to the ML literature, a move I thought was a mistake. However, looking at their published research or their 2017 review page, in practice this seems to have been less of a change of direction than I had thought, as most of their work appears to remain on highly differentiated and unreplaceable agent foundations type work - it seems unlikely that anyone not motivated by AI safety would produce this work. Even within those concerned about friendly AI, few not at MIRI would produce this work.

Critch's *Toward Negotiable Reinforcement Learning: Shifting Priorities in Pareto Optimal Sequential Decision-Making* (elsewhere titled '*Servant of Many Masters*') is a neat paper. Basically it identifies the pareto-efficient outcome if you have two agents with different beliefs who want to agree on a utility function for an AI, in a generalisation of Harsanyi's *Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility*. The key assumption is both want to use their current beliefs when they calculate the expected value of the deal to themselves, and the (surprising to me) conclusion is that over time the AI will have to weigh more and more heavily the *values* of the negotiator whose *beliefs* were more accurate. While I don't think this is necessarily Critch's interpretation, I take this as something of a *reductio* of the assumption. Surely if I was negotiating over a utility function, I would want the agent to learn about the world and use that knowledge to better promote my values ... not to learn about the world, decide I was a moron with a bad world model, and ignore me thereafter? If I think the AI is/will be smarter than me, I should be happy for it to do things I'm unaware will benefit me, and avoid doing things I falsely believe will help me. On the other hand, if the parties are well-informed nation states rather than individuals, the prospect of ‘getting one over’ the other might be helpful for avoiding arms races?

Kosoy's *Optimal polynomial-time estimators* addresses a similar topic to the Logical Induction work - assigning 'probabilities' to logical/mathematical/deductive statements

under computational limitations - but with a quite different approach to solving it. The work seems impressive but I didn't really understand it. Inside his framework he can prove that various results from probability theory also apply to logical statements, which seems like what we'd want. (Note that technically this paper came out in December 2016, and so is included in this year rather than last year's.)

Carey's article, *Incorrigibility in the CIRL Framework*, is a response to Milli et al.'s *Should Robots be Obedient* and Hadfield-Menel's *The Off-Switch Game*. Carey basically argues it's not necessarily the case that the CIRLs will be 'automatically' corigible if the AI's beliefs about value are very wrong, for example due to incorrect parameterisation or assigning a zero prior to something that turns out to be the case. The discussion section has some interesting arguments, for example pointing out that an algorithm designed to shut itself off unless it had a track record of perfectly predicting what humans would want might still fail if its ontology was insufficient, so it couldn't even tell that it was disagreeing with the humans during training. I agree that value complexity and fragility might mean it's very likely that any AI's value model will be partially (and hence, for an AGI, catastrophically) mis-parameterised. However, I'm not sure how much the examples that take up much of the paper add to this argument. Milli's argument only holds when the AI can learn the parameters, and given that this paper assumes the humans choose the wrong action by accident less than 1% of the time, it seems that the AI should assign a very large amount of evidence to a shutdown command... instead the AI seems to simply ignore it?

Some of MIRI's publications this year seem to mainly be better explanations of previous work. For example, Garrabrant et al's *A Formal Approach to the Problem of Logical Non-Omniscience* seems to be basically an easier to understand version of last year's *Logical Induction*. Likewise Yudkowsky and Soares's *Functional Decision Theory: A New Theory of Instrumental Rationality* seems to be basically new exposition of classic MIRI/LW decision theory work - see for example Soares et al's *Toward Idealized Decision Theory*. Similarly, I didn't feel like there was much new in Soares et al's *Cheating Death in Damascus*. Making things easier to understand is useful - and last year's Logical Induction paper was a little dense - but it's clearly not as impressive as inventing new things.

When I asked for top achievements for 2017, MIRI pointed me towards a lot of work they'd posted on agentfoundations.org as being one of their major achievements for the year, especially *this*, *this* and *this*, which pose and then solve a problem about how to find game-theoretic agents that can stably model each other, formulated it as a topological fixed point problem. There is also a lot of other work on agentfoundations that seems interesting, I'm not entirely sure how to think about giving credit for these. These seem more like 'work in progress' than finished work - for most organisations I am only giving credit for the latter. MIRI could with some justification respond that the standard academic process is very inefficient, and part of their reason for existence is to do things that universities cannot. However, even if you de-prioritise peer review, I still think it is important to write things up into papers. Otherwise it is extremely hard for outsiders to evaluate - bad both for potential funders and for people wishing to enter the field. Unfortunately it is possible that, if they continue on this route, MIRI might produce a lot of valuable work that is increasingly illegible from the outside. So overall I think I consider these as evidence that MIRI is continuing to actually do research, but will wait until they're ArXived to actually review them. If you disagree with this approach, MIRI is going to look much more productive, and their research possibility accelerating in 2017 vs 2016. If you instead only look at published papers, 2017 appears to be something of a 'down year' after 2016.

Last year I was not keen to see that Eliezer was spending a lot of time producing content on Arbital as part of his job at MIRI, as there was a clear conflict of interest - he was a significant shareholder in Arbital, and additionally I expected Arbital to fail. Now that Arbital does seem to have indeed failed, I'm pleased he seems to be spending less time on it, but confused why he is spending any time at all on it - though some of this seems to be cross-posted from elsewhere.

Eliezer's book *Inadequate Equilibria*, however, does seem to be high quality - basically another sequence - though only relevant inasmuch as AI safety might be one of many applications of the subject of the book. I also encourage readers to also read this excellent article by Greg Lewis (FHI) on the other side.

I also enjoyed *There's No Fire Alarm for Artificial General Intelligence*, which although accessible to the layman I think provided a convincing case that, even when AGI is imminent, there would (/might be) no signal that this was the case, and his socratic security dialogs on the mindset required to develop a secure AI.

I was sorry to hear Jessica Taylor left MIRI, as I thought she did good work.

MIRI spent roughly $1.9m in 2017, and aim to rapidly increase this to $3.5m in 2019, to fund new researchers and their new engineering team.

The Open Philanthropy Project awarded MIRI a $3.75m grant (over 3 years) earlier this year, largely because one reviewer was impressed with their work on Logical Induction. You may recall this was a significant part of why I endorsed MIRI last year. However, as this review is focused on work in the last twelve months, they don't get credit for the same work two years running! OPP have said they plan to fund roughly half of MIRI's budget. On the positive side, one might argue this was essentially a 1:1 match on donations to MIRI - but there are clearly game-theoretic problems here. Additionally, if you had faith in OpenPhil's process, you might consider this a positive signal of MIRI quality. On the other hand, if you think MIRI's marginal cost-effectiveness is diminishing over the multi-million dollar range, this might reduce your estimate of the cost-effectiveness of the marginal dollar.

There is also $1m of somewhat plausibly counterfactually valid donation matching available for MIRI (but not other AI Xrisk organisations).

Finally, I will note that MIRI are have been very generous with their time in helping me understand what they are doing.

# The Future of Humanity Institute (FHI)

Oxford's FHI requested not to be included in this analysis, so I won't be making any comment on whether or not they are a good place to fund. Had they not declined (and depending on their funding situation) they would have been a strong candidate. This was disappointing to me, because they seem to have produced an impressive list of publications this year, including a lot of collaborations. I'll briefly note two some pieces of research they published this year, but regret not being able to give them better coverage.

Saunders et al. published *Trial without Error: Towards Safe Reinforcement Learning via Human Intervention*, a nice paper where they attempt to make a Reinforcement Learner that can 'safely' learn by training a catastrophe-recognition algorithm to

oversee the training. It's a cute idea, and a nice use of the OpenAI Atari suite, though I was most impressed with the fact that they concluded that their approach would not scale (i.e. would not work). It's not often researchers publish negative results!

Honourable mention also goes to the very cool (but aren't all his papers?) Sandberg et al. *That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox*, which is relevant inasmuch as it suggests that the Fermi Paradox is *not* actually evidence against AI as an existential risk.

FHI's Brundage Bot apparently reads every ML paper ever written.

# Global Catastrophic Risks Institute (GCRI)

The Global Catastrophic Risks Institute is run by Seth Baum and Tony Barrett. They have produced work on a variety of existential risks, including non-AI risks. Some of this work seems quite valuable, especially Denkenberger's *Feeding Everyone No Matter What* on ensuring food supply in the event of disaster, and is probably probably of interest to the sort of person who would read this document. However, they are off-topic for us here. Within AI they do a lot of work on the strategic landscape, and are very prolific.

Baum's *Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy* attempts to analyse all existing AGI research projects. This is a huge project and I laud him for it. I don't know how much here is news to people who are very plugged in, but to me at least it was very informative. The one criticism I would have is it could do more to try to differentiate on capacity/credibility - e.g. my impression is Deepmind is dramatically more capable than many of the smaller organisations listed - but that is clearly a very difficult ask. It's hard for me to judge the accuracy, but I didn't notice any mistakes (beyond being surprised that AIXI has an 'unspecified' for safety engagement, given the amount of AI safety papers coming out of ANU.)

Baum's *Social Choice Ethics in Artificial Intelligence* argues that value-learning type approaches to AI ethics (like CEV ) contain many degrees of freedom for the programmers to finesse it to pick their values, making them no better than the programmers simply choosing an ethical system directly. The programmers can choose *whose* values are used for learning, how they are *measured*, and how they are *aggregated*. Overall I'm not fully convinced - for example, *pace* the argument on page 3, a Law of Large Numbers argument could support averaging many views to get at the true ethics *even if we had no way of independently verifying the true ethics*. And there is some irony that, for all the paper's concern with bias risk, the left-wing views of the author come through strongly. But despite these I liked the paper, especially for the discussion of who has standing - something that seems like it will need a philosophical solution, rather than a ML one.

Barrett's *Value of Global Catastrophic Risk (GCR) Information: Cost-Effectiveness-Based Approach for GCR Reduction* covers a lot of familiar ground, and then attempts to do some monte carlo cost-benefit analysis on the a small number of interventions to help address nuclear war and comet impact. After putting a lot of thought into setting up the machinery, it would have been good to see analysis of a wider range of risks!

Baum & Barrett published *Global Catastrophes: The Most Extreme Risks*, which seems to be essentially a reasonably well argued general introduction to the subject of

existential risks. Hopefully people who bought the book for other reasons will read it and become convinced.

Baum & Barrett's *Towards an Integrated Assessment of Global Catastrophic Risk* is a similar introductory piece on catastrophic risks, but the venue - a colloquium on catastrophic risks - seems less useful, as people reading it are more likely to already be concerned about the subject, and I don't think it spends enough time on AI risk *per se* to convince those who were already worried about Xrisk but not AI Xrisk.

Last year I was (and still am) impressed by their paper *On the Promotion of Safe and Socially Beneficial Artificial Intelligence*, which made insightful, convincing and actionable criticisms of 'AI arms race' language. I was less convinced by this year's *Reconciliation Between Factions Focused on Near-Term and Long-Term Artificial Intelligence*, which argues for a re-alignment away from near-term AI worries vs long-term AI worries towards AI worriers vs non-worriers. However, I'm not sure why anyone would agree to this - long-term worriers don't currently spend much time arguing against short-term worries (even if you thought that AI discrimination arguments were orwellian, why bother arguing about it?), and convincing short-term worriers to stop criticise long-term worries seems approximately as hard as simply convincing them to become long-term worriers.

GCRI spent approximately $117k in 2017, which is shockingly low considering their productivity. This was lower than 2016; apparently their grants from the US Dept. of Homeland Security came to an end.

# The Center for the Study of Existential Risk (CSER)

CSER is an existential risk focused group located in Cambridge. Like GCRI they do work on a variety of issues, notably including Rees' work on infrastructure resilience.

Last year I criticised them for not having produced any online research over several years; they now have a separate page that does list some but maybe not all of their research.

Liu, a CSER researcher, wrote *The Sure-Thing principle and P2* and was second author on Gaifman & Liu's *A simpler and more realistic subjective decision theory*, both on the mathematical foundations of bayesian decision theory, which is a valuable topic for AI safety in general. Strangely neither paper mentioned CSER as a financial supporter of the paper or affiliation.

Liu and Price's *Heart of DARCness* argues that agents do not have credences for what they will do while deciding whether to do it - their confidence is temporarily undefined. I was not convinced - even someone is deciding whether she's 75% confident or 50% confident, presumably there are some odds that determine which side in a bet she'd take if forced to choose? I'm also not sure of the direct link to AI safety.

They've also convened and attended workshops on AI and decision theory, notably the AI & Society Symposium in Japan, but in general I am wary of giving organisations credit for these, as they are too hard for the outside observer to judge, and ideally workshops lead to produce papers - in which case we can judge those.

CSER also did a significant amount of outreach, including presenting to the House of Lords, and apparently have expertise in Chinese outreach (multiple native mandarin speakers), which could be important, given China's AI research but cultural separation from the west.

They are undertaking a novel publicity effort that I won't name as I'm not sure it's public yet. In general I think most paths to success involve consensus-building among mainstream ML researchers, and 'popular' efforts risk harming our credibility, so I am not optimistic here.

Their annual budget is around $750,000, with I estimate a bit less than half going on AI risk . Apparently they need to raise funds to continue existing once their current grants run out in 2019.

# AI Impacts

AI Impacts is a small group that does high-level strategy work, especially on AI timelines, somewhat associated with MIRI.

They seem to have produced significantly more this year than last year. The main achievement is the *When will AI exceed Human Performance? Evidence from AI Experts*, which asked gathered the opinions of hundreds of AI researchers on AI timelines questions. There were some pretty relevant takeaways, like that most researchers find the AI Catastrophic Risk argument somewhat plausible, but doubt there is anything that can usefully be done in the short term, or that asian researchers think human-level AI is significantly closer than americans do. I think the value-prop here is twofold: firstly, providing a source of timeline estimates for when we make decisions that hinge on how long we have, and secondly, to prove that concern about AI risk is a respectable, mainstream position. It was apparently one of the most discussed papers of 2017.

On a similar note they also have data on improvements in a number of AI-related benchmarks, like computing costs or algorithmic progress.

John Salvatier (member of AI Impacts at the time) was also second author on Agent-Agnostic Human-in-the-Loop Reinforcement Learning, along with Evans (FHI, 4th author), which attempts to design an interface for reinforcement learning that abstracts away from the agent, so you could easily change the underlying agent.

AI Impacts' budget is tiny compared to most of the other organisations listed here; around $60k at present. Incremental funds would apparently be spent on hiring more part-time researchers.

# Center for Human-Compatible AI (CFHCA)

The Center for Human-Compatible AI, founded by Stuart Russell in Berkeley, launched in August 2016. As they are not looking for more funding at the moment I will only briefly survey some of they work on cooperative inverse reinforcement learning.

Hadfield-Menel et al's *The Off-Switch Game* is a nice paper that produces and formalises the (at least now I've read it) very intuitive result that a value-learning AI

might be corrigible (at least in some instances) because it takes the fact that a human pressed the off-switch as evidence that this is the best thing to do.

Milli et al's *Should Robots be Obedient* is in the same vein as Hadfield-Menel et al's *Cooperative Inverse Reinforcement Learning* (last year) on learning values from humans, specifically touching on whether such agents would be willing to obey a command to 'turn off', as per Soares's paper on *Corrigibility*. She does some interesting analysis about the trade-off between obedience and results in cases where humans are fallible.

In both cases I thought the papers were thoughtful and had good analysis. However, I don't think either is convincing in showing that corrigibility comes 'naturally' - at least not the strength of corrigibility we need.

I encourage them to keep their website more up-to-date.

Overall I think their research is good and their team promising. However, apparently they have enough funding for now, so I won't be donating this year. If this changed and they requested incremental capital I could certainly imagine funding them in future years.

# Other related organisations

The Center for Applied Rationality (CFAR) works on trying to improve human rationality, especially with the aim of helping with AI Xrisk efforts.

The Future of Life Institute (FLI) ran a huge grant-making program to try to seed the field of AI safety research. There definitely seem to be a lot more academics working on the problem now, but it's hard to tell how much to attribute to FLI.

Eighty Thousand Hours (80K) provide career advice, with AI safety being one of their key cause areas.

# Related Work by other parties

*Deep Reinforcement Learning from Human Preferences*, was possibly my favourite paper of the year, which possibly shouldn't come as a surprise, given that two of the authors (Christiano and Amodei from OpenAI ) were authors on last year's *Concrete Problems in AI Safety*. It applies ideas on bootstrapping that Christiano has been discussing for a while - getting humans to train an AI which then trains another AI etc. The model performs significantly better than I would have expected, and as ever I'm pleased to see OpenAI - Deepmind collaboration.

Christiano continues to produce very interesting content on his blog, like this on Corrigibility. When I first read his articles about how to bootstrap safety through iterative training procedures, my reactions was that, while this seemed an interesting idea, it didn't seem to have much in common with mainstream ML. However, there do seem to be a bunch of practical papers about imitation learning now. I'm not sure if this was always the case, and I was just ignorant, or if they have become more prominent in the last year. Either way, I have updated towards considering this approach to be a promising one for integrating safety into mainstream ML work. He

has also written a nice blog post explaining how AlphaZero works, and arguing that this supports his enhancement ideas.

It was also nice to see ~95 papers that were addressing Amodei et al's call in last year's *Concrete Problems*.

Menda et al's *DropoutDAgger* paper on safe exploration seems to fit in this category. Basically they come up with a form of imitation learning where the AI being trained can explore a bit, but isn't allowed to stray too far from the expert policy - though I'm not sure why they always have the learner explore in the direction it thinks is best, rather than assigning some weight to its uncertainty of outcome, explore-exploit-style. I'm not sure how much credit Amodei et al can get for inspiring this though, as it seems to be (to a significant degree) an extension of Zhang and Cho's *Query-Efficient Imitation Learning for End-to-End Autonomous Driving*.

However, I don't want to give too much credit for work that improves 'local' safety that doesn't also address the big problems in AI safety, because this work probably accelerates unsafe human-level AI. There are many papers in this category, but for obvious reasons I won't call them out.

Gan's *Self-Regulating Artificial General Intelligence* contains some nice economic formalism around AIs seizing power from humans, and raises the interesting argument that if you need specialist AIs to achieve things, the first human-level AIs might not exhibit takeoff behaviour because they would be unable to sufficiently trust the power-seizing agents they would need to create. I'm sceptical that this assumption about the need for specialised AIs holds - surely even if you need to make separate AI agents for different tasks, rather than integrating them, it would suffice to give them specialised *capabilities* and but the same *goals*. Regardless, the paper does suggest the interesting possibility that humanity might make an AI which is intelligent enough to realise it cannot solve the alignment problem to safely self-improve... and hence progress stops there - though of course this would not be something to rely on.

MacFie's *Plausibility and Probability in Deductive Reasoning* also addresses the issue of how to assign probabilities to logical statements, in a similar vein to much MIRI research.

Vamplew et al's *Human-aligned artificial intelligence is a multiobjective problem* argues that we should consider a broader class of functions than linear sums when combining utility functions.

Google Deepmind continue to churn out impressive research, some of which seems relevant to the problem, like Sunehag et al's *Value-Decomposition Networks For Cooperative Multi-Agent Learning* and Danihelka, et al's *Comparison of Maximum Likelihood and GAN-based training of Real NVPs* on avoiding overfitting.

In terms of predicting AI timelines, another piece I found interesting was Gupta et al.'s *Revisiting the Unreasonable Effectiveness of Data*, which argued that, for vision tasks at least, performance improved logarithmically in sample size.

The Foresight Institute published a white paper on the general subject of AI policy and risk.

Stanford's *One Hundred Year Study on Artificial Intelligence* produced an AI Index report, which is basically a report on progress in the field up to 2016. Interestingly various metrics they tracked, summarised in their 'Vibrancy' metric, suggest that the

field actually regressed in 2016, through my experience with similar data in the financial world leaves me rather sceptical of such methodology. Unfortunately the report dedicated only a single word to the subject of AI safety.

On a lighter note, the esteemed G.K. Chesterton returned from beyond the grave to eviscerate an AI risk doubter, and a group of researchers (some FHI) *proved* that it is impossible to create a machine larger than a human, so that's a relief.

# Other major developments this year

Google's Deepmind produced AlphaZero, which learnt how to beat the best AIs (and hence also the best humans) at Go, Chess and Shogi with just a few hours of self-play.

Creation of the EA funds, including the *Long-Term Future Fund*, run by Nick Beckstead, which has made one smallish grant related to AI Safety, conserved the other 96%.

The Open Philanthropy Project funded both MIRI and OpenAI (acquiring a board seat in the process with the latter).

Nvidia (who make GPUs used for ML) saw their share price approximately doubl, after quadrupling last year.

Hillary Clinton was possibly concerned about AI risk? But unfortunately Putin seems to have less helpful concerns about an AI Arms race... namely ensuring that *he wins it*. And China announced a national plan for AI with chinese characteristics - but bear in mind they have failed at these before, like their push into Semiconductors, though companies like Baidu do seem to be doing impressive research.

There were some papers suggesting the replication crisis may be coming to ML?

# Conclusion

In some ways this has been a great year. My impression is that the cause of AI safety has become increasingly mainstream, with a lot of researchers unaffiliated with the above organisations working at least tangentially on it.

However, it's tough from the point of view of an external donor. Some of the organisations doing the best work are well funded. Others (MIRI) seem to be doing a lot of good work but (perhaps necessarily) it is significantly harder for outsiders to judge than last year, as there doesn't seem to be a really heavy-hitting paper like there was last year. I see MIRI's work as being a long-shot bet that their specific view of the strategic landscape is correct, but given this they're basically irreplaceable. GCRI and CSER's work is more mainstream in this regard, but GCRI's productivity is especially noteworthy, given the order of magnitude of difference in budget size.

As I have once again failed to reduce charity selection to a science, I've instead attempted to subjectively weigh the productivity of the different organisations against the resources they used to generate that output, and donate accordingly.

My constant wish is to promote a lively intellect and independent decision-making among my readers; hopefully my laying out the facts as I see them above will prove

helpful to some readers. Here is my eventual decision, rot13'd so you can do come to your own conclusions first if you wish:

*Fvtavsvpnag qbangvbaf gb gur Znpuvar Vagryyvtrapr Erfrnepu Vafgvghgr naq gur Tybony Pngnfgebcuvp Evfxf Vafgvghgr. N zhpu fznyyre bar gb NV Vzcnpgf.*

However I wish to emphasis that all the above organisations seem to be doing good work on the most important issue facing mankind. It is the nature of making decisions under scarcity that we must prioritize some over others, and I hope that all organisations will understand that this necessarily involves negative comparisons at times.

Thanks for reading this far; hopefully you found it useful. Someone suggested that, instead of doing this annually, I should instead make a blog where I provide some analysis of AI-risk related events as they occur. Presumably there would still be an annual giving-season writeup like this one. If you'd find this useful, please let me know.

# Disclosures

I was a Summer Fellow at MIRI back when it was SIAI, volunteered very briefly at GWWC (part of CEA) and once applied for a job at FHI. I am personal friends with people at MIRI, FHI, CSER, CFHCA and AI Impacts *but not GCRI* (so if you're worried about bias you should overweight them… though it also means I have less direct knowledge). However I have no financial ties beyond being a donor and have never been romantically involved with anyone who has ever been at any of the organisations.

I shared a draft of the relevant sections of this document with representatives of MIRI, CSER and GCRI and AI Impacts. I'm very grateful for Alex Flint and Jess Riedel for helping review a draft of this document. Any remaining inadequacies and mistakes are my own.

*Edited 2017-12-21: Spelling mistakes, corrected Amodei's affiliation.*

*Edited 2017-12-24: Minor correction to CSER numbers.*

# Bibliography


Adam D. Cobb, Andrew Markham, Stephen J. Roberts; Learning from lions: inferring the utility of agents from their trajectories; https://arxiv.org/abs/1709.02357
Alexei Andreev; What's up with Arbital; http://lesswrong.com/r/discussion/lw/otq/whats_up_with_arbital/
Allison Duettmann; Artificial General Intelligence: Timeframes & Policy White Paper; https://foresight.org/publications/AGI-Timeframes&PolicyWhitePaper.pdf
Anders Sandberg, Stuart Armstrong, Milan Cirkovic; That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox; https://arxiv.org/pdf/1705.03394.pdf
Andrew Critch, Stuart Russell; Servant of Many Masters: Shifting priorities in Pareto-optimal sequential decision-making; https://arxiv.org/abs/1711.00363
Andrew Critch; Toward Negotible Reinforcement Learning: Shifting Priorities in Pareto Optimal Sequential Decision-Making; https://arxiv.org/abs/1701.01302

Andrew MacFie; Plausibility and Probability in Deductive Reasoning; https://arxiv.org/pdf/1708.09032.pdf
Assaf Arbelle, Tammy Riklin Raviv; Microscopy Cell Segmentation via Adversarial Neural Networks; https://arxiv.org/abs/1709.05860
Ben Garfinkel, Miles Brundage, Daniel Filan, Carrick Flynn, Jelena Luketina, Michael Page, Anders Sandberg, Andrew Snyder-Beattie, and Max Tegmark; On the Impossibility of Supersized Machines; https://arxiv.org/pdf/1703.10987.pdf
Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, Sergey Levine; One-Shot Visual Imitation Learning via Meta-Learning; https://arxiv.org/abs/1709.04905
Chen Sun, Abhinav Shrivastava Saurabh Singh, Abhinav Gupta; Revisiting Unreasonable Effectiveness of Data in Deep Learning Era; https://arxiv.org/pdf/1707.02968.pdf
Chih-Hong Cheng, Frederik Diehl, Yassine Hamza, Gereon Hinz, Georg Nuhrenberg, Markus Rickert, Harald Ruess, Michael Troung-Le; Neural Networks for Safety-Critical Applications - Challenges, Experiments and Perspectives; https://arxiv.org/pdf/1709.00911.pdf
Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané; Concrete Problems in AI Safety; https://arxiv.org/abs/1606.06565
David Abel, John Salvatier, Andreas Stuhlmüller, Owain Evans; Agent-Agnostic Human-in-the-Loop Reinforcement Learning; https://arxiv.org/abs/1701.04079
Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell; The Off-Switch Game; https://arxiv.org/pdf/1611.08219.pdf
Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell; Cooperative Inverse Reinforcement Learning; https://arxiv.org/abs/1606.03137
Eliezer Yudkowsky and Nate Soares; Functional Decision Theory: A New Theory of Instrumental Rationality; https://arxiv.org/abs/1710.05060
Eliezer Yudkowsky; A reply to Francois Chollet on intelligence exposion; https://intelligence.org/2017/12/06/chollet/
Eliezer Yudkowsky; Coherant Extrapolated Volition; https://intelligence.org/files/CEV.pdf
Eliezer Yudkowsky; Inadequate Equilibria; https://www.amazon.com/dp/B076Z64CPG
Eliezer Yudkowsky; There's No Fire Alarm for Artificial General Intelligence; https://intelligence.org/2017/10/13/fire-alarm/
Filipe Rodrigues, Francisco Pereira; Deep learning from crowds; https://arxiv.org/abs/1709.01779
Greg Lewis; In Defense of Epistemic Modesty; http://effective-altruism.com/ea/1g7/in_defence_of_epistemic_modesty/
Haim Gaifman and Yang Liu; A simpler and more realistic subjective decision theory; https://link.springer.com/article/10.1007%2Fs11229-017-1594-6
Harsanyi; Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility; http://www.springer.com/us/book/9789027711861
Ivo Danihelka, Balaji Lakshminarayanan, Benigno Uria, Daan Wierstra, Peter Dayan; Comparison of Maximum Likelihood and GAN-based training of Real NVPs; https://arxiv.org/pdf/1705.05263.pdf
Jiakai Zhang, Kyunghyun Cho; Query-Efficient Imitation Learning for End-to-End Autonomous Driving; https://arxiv.org/abs/1605.06450
Joshua Gans; Self-Regulating Artificial General Intelligence; https://arxiv.org/pdf/1711.04309.pdf
Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, Owain Evans; When will AI exceed Human Performance? Evidence from AI Experts; https://arxiv.org/abs/1705.08807
Kavosh Asadi, Cameron Allen, Melrose Roderick, Abdel-rahman Mohamed, George Konidaris, Michael Littman; Mean Actor Critic; https://arxiv.org/abs/1709.00503
Kunal Menda, Katherine Driggs-Campbell, Mykel J. Kochenderfer; DropoutDAgger: A Bayesian Approach to Safe Imitation Learning; https://arxiv.org/abs/1709.06166

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, Olivier Bousquet; Are GANs Created Equal? A Large-Scale Study; https://arxiv.org/abs/1711.10337
Martin Rees; "Black Sky" Infrastructure and Societal Resilience Workshop; https://www.cser.ac.uk/media/uploads/files/Black-Sky-Workshop-at-the-Royal-Society-Jan.-20171.pdf
Mile Brundage; Brundage Bot; https://twitter.com/BrundageBot
Minghai Qin, Chao Sun, Dejan Vucinic; Robustness of Neural Networks against Storage Media Errors; https://arxiv.org/abs/1709.06173
Myself; 2017 AI Risk Literature Review and Charity Evaluation; http://effective-altruism.com/ea/14w/2017_ai_risk_literature_review_and_charity/
Nate Soares and Benja Fallenstein; Towards Idealized Decision Theory; https://arxiv.org/pdf/1507.01986.pdf
Nate Soares and Benjamin Levinstein; Cheating Death in Damascus; https://intelligence.org/files/DeathInDamascus.pdf
Nates Soares, Benja Fallenstein, Eliezer Yudkowsky, Stuart Armstrong; Corrigibility; https://intelligence.org/files/Corrigibility.pdf
Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, Dario Amodei; Deep Reinforcement Learning from Human Preferences; https://arxiv.org/abs/1706.03741
Paul Christiano; AlphaGo Zero and capability amplification; https://ai-alignment.com/alphago-zero-and-capability-amplification-ede767bb8446
Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, David Meger; Deep Reinforcement Learning that Matters; https://arxiv.org/abs/1709.06560
Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, Astro Teller.; One Hundred Year Study on Artificial Intelligence; https://ai100.stanford.edu/
Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, Thore Graepel; Value-Decomposition Networks For Cooperative Multi-Agent Learning; https://arxiv.org/pdf/1706.05296.pdf
Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, Jane Mummery; Human-aligned artificial intelligence is a multiobjective problem; https://link.springer.com/article/10.1007/s10676-017-9440-6
Ryan Carey; Incorrigibility in the CIRL Framework; https://arxiv.org/abs/1709.06275
Samuel Yeom, Matt Fredrikson, Somesh Jha; The Unintended Consequences of Overfitting: Training Data Inference Attacks; https://arxiv.org/abs/1709.01604
Scott Alexander; G.K. Chesterton on AI Risk; http://slatestarcodex.com/2017/04/01/g-k-chesterton-on-ai-risk/
Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, Jessica Taylor; A Formal Approach to the Problem of Logical Non-Omniscience; https://arxiv.org/abs/1707.08747
Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, Jessica Taylor; Logical Induction; http://arxiv.org/abs/1609.03543
Seth Baum and Tony Barrett; Global Catastrophes: The Most Extreme Risks; http://sethbaum.com/ac/2018_Extreme.pdf
Seth Baum and Tony Barrett; Towards an Integrated Assessment of Global Catastrophic Risk ; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046816
Seth Baum; On the Promotion of Safe and Socially Beneficial Artificial Intelligence; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2816323
Seth Baum; Reconciliation Between Factions Focused on Near-Term and Long-Term Artificial Intelligence; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2976444
Seth Baum; Social Choice Ethics in Artificial Intelligence;

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046725
Seth Baum; Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741
Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, Stuart Russell; Should Robots be Obedient; https://arxiv.org/pdf/1705.09990.pdf
Tony Barrett; Value of Global Catastrophic Risk (GCR) Information: Cost-Effectiveness-Based Approach for GCR Reduction; https://www.dropbox.com/s/7a7eh2law7tbvk0/2017-barrett.pdf?dl=0
Vadim Kosoy; Optimal Polynomial-Time Estimators: A Bayesian Notion of Approximation Algorithm; https://arxiv.org/abs/1608.04112
Victor Shih, David C Jangraw, Paul Sajda, Sameer Saproo; Towards personalized human AI interaction - adapting the behavior of AI agents using neural signatures of subjective interest; https://arxiv.org/abs/1709.04574
William Saunders, Girish Sastry, Andreas Stuhlmueller, Owain Evans; Trial without Error: Towards Safe Reinforcement Learning via Human Intervention; https://arxiv.org/abs/1707.05173
Xiongzhao Wang, Varuna De Silva, Ahmet Kondoz; Agent-based Learning for Driving Policy Learning in Connected and Autonomous Vehicles; https://arxiv.org/abs/1709.04622
Yang Liu and Huw Price; Heart of DARCness; http://yliu.net/wp-content/uploads/darcness.pdf
Yang Liu; The Sure-Thing principle and P2; http://www.academia.edu/33992500/The_Sure-thing_Principle_and_P2
Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, Byron Boots; Agile Off-Road Autonomous Driving Using End-to-End Deep Imitation Learning; https://arxiv.org/abs/1709.07174

# The Basic Object Model and Definition by Interface

What does it mean to say that something "exists"? Why do we say that both objects in the world, such as chairs, and logical entities, such as the number 9, both exist? I believe that the key insight is that both "things" can be adapted to what I call the Basic Object Model. Further, by using this specific example, I can demonstrate what I call Definition by Interface. This post responds to Philosopher Corner: Numbers.

**Update**: After further consideration, I have come to the conclusion that this is only part of the story. Fitting the object modle is a key part of "existence", but we also need to have a divide between "existence" and "non-existence" to fully explain why we are tempted to use the same term for both kinds of "objects". I hope to develop this in a future post.

### The Basic Object Model

Informally: any "collection" of "things" is adapted to the object model if we can say the following:

- Any "thing" in this "collection" has properties. For example, a table may have the properties of size, number of legs or general color. Number may have the properties of being odd or even, postitive or negative, prime or composite.
- There exist relations between the "things" in the "collection". For example, a table may be larger or smaller than another, to the left or to the right of a couch or darker or lighter than the blinds. Numbers may be smaller or larger, have more factors or less factors, have the same sign or different signs.
- Any "thing" has a type and all objects of this type will have certain properties and relations. For example, all tables have a size and a bigger than/smaller than relation to other tables. Some objects of the same type may have additional relations.
- If a "thing" X and a "thing" Y have the same identity, then they are the same type and all properties and relations are the same. So if X and Y are tables with the same identity, then the have the same size, number of legs and general color; as well as having the same size relation and positional relation to a couch.

"Exists" is a lingustic construct and a major factor in lingustic constructs is convenience. Since both physical objects and logical objects fit the Basic Object Model it is convenient to say that they both "exist". On the other hand, we tend not to say that chair's brownness or the number two's evenness 'exists', because it is usually more convenient to simply conceptualise them using a Property Inferface, rather than the Object Interface. To be clear, I'm not claiming that this is anyone's explicit reason, just that the presence of these similarities nudged us towards using the same word for both kinds of objects.

Further, I've only clarified what Existence in the Broad Sense means. Physical existence is a narrower kind of existence and there is more to that than just meeting the Basic Object Model. Similarly, it seems that there could be more to logical existence as well, or if not, perhaps maths specifically exists in a deeper sense. What I mean here is that if we performed a conceptual analysis on Object Existence and a conceptual analysis on Mathematical Existence, I would expect that we would find that

the linguistic term conveys more than just the Basic Object Model when used in these narrow senses. Further, we would find that these more specific uses would differ with the Basic Object Model being most or perhaps even all of what they have in common.

As I said at the start, the Basic Object Model is an example of what I call Definition by Interface. Terms that are defined in this way can simply be adapted to the same interface, instead of necessarily referring to some distinction that actually exists in the ontology. Some of these terms may be Defined by Interface and also exist in the ontology, but we should not assume such a deeper existance without good reason. In particular, just because a word seems as fundamental as "existence", we should not fall into the trap of assuming that it must be ontologically basic. And as we've seen here, even though its use in specific cases may be ontologically basic if we dig into them, the word in general just seems to refer to similarities in the interface.

# Truth is Symmetric

The Taj Mahal is symmetric with respect to right-left reflection. That is to say, if you create a mirror image of it that flips right and left, you get the exact same thing. Therefore, if as we're standing outside looking at it my friend Jen tells me that she thinks the left side of the Taj Mahal is much nicer than the right side, or that it's objectively better, I would be suspicious. That's because if we created the mirror image of the Taj Mahal, the right side would be on the left, Jen would not be able to tell the difference and so would say that the left side is nicer than the right one, but it would secretly be the right side she was speaking about. Muwahahahaa, it can't be right. At least if all that determines niceness isn't about where something is placed but what it's composed of - the left side can be nicer than the right, but not just because it's on the left.

Now for three similar examples, each trivial on its own but illustrating a broader point.

1) Most Israeli Jews think Israel is broadly in the right in the Israeli-Palsetinian conflict. Most Palestinians think that the Palestinians are broadly in the right. What bothers me about this is that both factions think that their faction is *objectively* in the right, that is, independently of who considers the question. But if it's independently of who considers the question, how come people who were born on one side of the aisle mostly think one thing, and people on the other side think another thing. This screams inconsistency and irrationality, since a factor that shouldn't affect what's objectively right affects what most people think.

2) I'm usually more left-wing in my opinions. So when I consider whether some extreme right-wing statement falls within the limits of free speech, I try to think whether I would consider an equally-extreme *left*-wing statement to fall within the limits of free speech. I imagine a world where that opinion is banned from being expressed, and whether I'd be happy with a world like that. If for the right-wing statement my feeling is "This statement should be banned!" but for the left-wing statement I feel "It would be a terrible world if that statement were banned" then I suspect myself of being biased, since a factor that shouldn't affect what falls within free speech (the alignment of the statement) correlates with my judgment.

3) There's a cliché that everyone driving faster than you is crazy but everyone driving slower than you is an idiot. Since being crazy and being an idiot are traits that are independent of who makes the judgment and many people make different judgments, you should suspect bias, since a factor that shouldn't affect whether someone is objectively an idiot (how fast *you* are driving) correlates with your belief.

(I'm not sure that craziness and idiocy are really objective. Rather, from inside they feel like they are, and can be defined objectively to accord to that - e.g. if the risk of that person dying outweighs their gain from arriving faster, they are "crazy" drivers.)

These are all examples of cases where I use symmetry to flag beliefs suspected of being biased. This is symmetry in the mathematical sense, of changing some factor which the result should not depend on (nationality determined by birth, political alignment of statement, speed you prefer driving) changes the belief about the result (objective moral judgment, free speech application, insanity/idiocy), and so that belief is suspicious. The important property here is that the result should be *invariant* to changing something, it's *symmetrical* with respect to that factor, the same way that

the Taj Mahal (and niceness) is symmetric with respect to reflecting it about the middle which replaces the right and left sides.

Before I go on to how I use it, I want to impress upon you that this line of argumentation is *very* common. Here are examples from SlateStarCodex which is very fond of it, though maybe implicitly. "The control group is out of control" uses the fact that we should judge scientific studies based on their methodology and not the subject matter - our scientific judgment should be symmetric with respect to changing the specific matter studied - and so if parapsychology has the same methodologically-robust results as other fields then if we believe results in other fields we should believe results in parapsychology (and then goes on to argue that since these results are probably false we should be suspicious of the other equally robust scientific results, to preserve the symmetry). "Guided by the beauty of our weapons" advocates for persuasion methods that would lead us towards the truth regardless of whether we are right or wrong when we first use them - weapons whose effects are symmetric with respect to the view of the wielder (which are confusingly called asymmetric in the post). "Against Murderism" shows many arguments that are symmetric with respect to substituting racism for "murderism", and argues that since with murderism we see these are absurd, we should reject them for racism as well.

Lastly, from "Proving too much":

> Proving Too Much is when you challenge an argument because, in addition to proving its intended conclusion, it also proves obviously false conclusions. For example, if someone says “You can’t be an atheist, because it’s impossible to disprove the existence of God”, you can answer “That argument proves too much. If we accept it, we must also accept that you can’t disbelieve in Bigfoot, since it’s impossible to disprove his existence as well.”

The argument might work for God, but then it would also work for Bigfoot as well - it's symmetric with respect to replacing God with Bigfoot, since it doesn't use any property which distinguishes them - and so if it proves something about God, it also proves the same thing for Bigfoot.

How I use it

Finding these irrelevant factors that affect our beliefs really helps in raising suspicion of them and not letting them fly under the radar - once you consider, many of these are easily exposed - but we often don't consider. I try to always be aware of irrelevant factors ("symmetries") that shouldn't affect the truth but that might affect my judgment on some matter, and then I try to simulate to see if I would have judged differently if those factors were different, and if so, flag that belief for inspection. For example, whenever I try to judge if free speech applies to some opinion, I flip that opinion about the center and see if it would apply to an equally-extreme opinion on the other side. When I try to form a judgment about tax policy, I try to think what I would have thought about it if I was poorer. Examples of these factors are all the biases we all know and love - sunk costs, cognitive dissonance, the affect heuristic, status-quo bias etc., but also factors that need not have a standard name but apply to a certain situation. There is no "the speed you prefer driving affects your judgment bias", though I'm sure it can be an instance of some bias.

Finding these symmetries is not easy (if it was, all these SSC posts linked above would have been self-evident, which they aren't), but it's not always very hard if you explicitly try. It's especially hard to find ones you can get good mileage out of, ones

that really often affect your judgment but really are symmetries for the outcome. For example, translating the time of judgment by one day and asking "would I still think the same thing tomorrow?" is a symmetry for many judgments but might not help you change your mind since it might not affect your judgment very often - you're as likely to make the same mistake about crazy drivers tomorrow (though it might help with present bias).

And if you find good symmetries for a specific issue you can use them whenever that issue comes up, so they're reusable. I use my symmetry about free speech sometimes, and the one that asks what I would have thought of a law as a person with the opposite interests very often. What are yours?

# Learning AI if you suck at math

This is a linkpost for https://hackernoon.com/learning-ai-if-you-suck-at-math-8bdfb4b79037

# Melting Gold, and Organizational Capacity

*Epistemic Status: Confident about the problem, solution and exact mechanism is more of an untested hypothesis.*

There's a saying for communities: if you're not gaining members, you're losing members. Sometimes you hit *just* the right size and you'd prefer things stay exactly as they are. But in practice, some people will eventually drift away - moving to a new town, getting a new job that consumes more time, or just losing interest in whatever-your-thing is.

So you need to have *some* source of new members, that supplies the life blood a community needs to keep going.

I think there's a similar truism:

*If you are not building organizational capacity, you are burning organizational capacity.*

(If the above sounds totally thoroughly reasonable and prompts an obvious course of action you can stop reading here, or skip towards the final section. Most of the rest of the post is for making this more salient)

Young organizations, communities and other groups tend to run on *hero power* - a couple people who care a lot who put in most of the work to keep it going. This is both because only a few people *do* care enough to put in that work, and because trying to distribute tasks is often harder than doing the tasks yourself.

As long those organizers are sticking around, everything seems fine. But the system is fragile.

I think people loosely understand the fragility. But I think it's better to think of it as a *resource being slowly burned down* rather than "a system is working fine until something changes."

## Nothing Gold Can Stay - and Gold Melts Suddenly.

At the end of the Lord of the Rings movies, a golden ring is dropped into Mount Doom. When I first watched the film, I expected the gold to slowly melt as it slipped into the magma. Instead, it stayed solidly ring shaped... until it abruptly melted in seconds.

*(This was the Ring of Power, approximately 3 seconds before it wasn't anymore)*

I thought that looked unrealistic... but someone told me this is how actual gold melts: temperature rising until suddenly it reaches a critical state change. (No comment on magical gold you have to take to a magic volcano)

Any time you have a mission critical job that *depends on a person being invested for it to get done*, I think you have something similar going on. (i.e. a job where a person isn't getting paid enough for it to be something they sustainably do as part of their general survival, especially if it requires them to exert a lot of agency or continuous attention).

From the outside, it looks like things are fine up until the organizer suddenly burns out. This may even be how it feels to the organizer ("the first stage of burnout is zeal, and the second stage of burnout is burnout"). But under the hood, some combination of thing are often happening:

1. They're slowly changing as a person, which will eventually make them no-longer-the-sort-of-person-who-wants-to-do-the-thing. They get bored of the thing, or they get interested in new things, or they just need a break.
2. A slow frustration/resentment eventually builds up that they doing the job without enough help.
3. They get physically worn out by stresses caused by the thing.

If you notice this far enough in advance, you can train new people to replace you. But this means you need either new heroes willing to do more work than they're paid for, or you need a whole lot of smaller-helpers who are somehow organized enough to do what had previously been a cohesive, unified job.

You may not find those people in time to replace you. Or, they may end up in the same dynamic you were in, and eventually abruptly burn out, and *then* may not be able to find people to replace *them.*

Or there may be a multi-step breakdown - you find someone who can do most of the things, but don't quite understand *all the pieces*, and then when they find a new person to replace *them*, they find a person who is able to mostly-do the pieces *they* understand well, but the pieces they understand less well get lost in second step of translation.

# Homemade Things Cost More

It costs more to build something yourself than to buy it factory made.

Things you make yourself are often able to be more unique and special than things mass-produced by capitalism. They can cater to special, niche interests without enough demand to *develop* mass production.

But there was a weird followup step, where Capitalism noticed that people had noticed that homemade things took more time and were worth more. And enterprising entrepreneurs saw free money and learned to *market* "homemade" things for more money.

As a result, I came to associate "homemade" with "overpriced." Many homemade things *aren't* that special and unique. An artisinal hand-crafted coffee mug isn't really worth more than a mass produced version on Amazon.

(Maybe this is where Premium Mediocre things come from?)

But... when the homemade thing *is* unique, when you literally can't get it anywhere else, and you are getting important social or cultural value from it... then... well, if you want that thing, the only way to get it is to pay homemade prices for it.

The problem is *you may not be able to pay for them with money.* They are usually labors of love. If there was *enough* demand for them for someone to do them full-time, you'd probably be able to mass produce them more cheaply anyway.

It's unlikely the people making them could actually *more easily* produce them if they were paid more. Or, the amount of money would be *dramatically* more than what seems obvious. It's not enough to cover costs. It has to be *enough to quit your day job,* and then *not worry about quitting your day job turning out to be a horrible idea.*

This means if you want to pay for a rare, precious thing that you want to keep existing, it is quite likely that the only ways to guarantee it's continued existence is to put in sweat and sacrifice. If things are well organized it shouldn't need to be a *major* sacrifice, but it may mean serious time and attention that you were spending on other things you cared about too.

I don't mean to say any of this in a *moralizing way.* This is not an essay about what you "should" do. This is just a description of what is in fact necessary for certain things to happen, if they are things that matter to you.

# Solution Hypotheses

My advice on what to do about this isn't really tested, but seem like the obvious things to consider:

**For Organizers -** Your job is not to *Do The Thing*. Your job is to make sure *The Thing Keeps Getting Done Whether Or Not You Do It.*

At first, it may seem nice and high-status to get all the credit for doing the thing. That credit will not sustain you forever - eventually you will probably need help, and it may happen more suddenly than you imagine.

My more speculative conjecture is: *as early as possible,* no matter *what* your job is, you should make a part of your job to find new people to start sharing the load.

As early as possible, you should also start investing in systems that make the job *easier.* Identify wasted motion. And ultimately streamline the *onboarding process* so new people have an easier time contributing.

This runs against my intuitions because doing the stuff myself is way easier than training a new person to do it. But I think it's important to consider this an essential skill, no matter what your task is. (Even if the primary task isn't very people-centric, you will need to develop the people-skills to identify suitable replacements and train them)

**For the People Enjoying The Thing** - If you are participating in a thing that is dependent on a few people putting in a herculean amount of effort...

Again, saying this without any intended moralizing, simply as a statement of fact: people can't run on respect and credit forever. And the situation often *can't be solved* simply by throwing money at it. (Or the amount of money is something like $40,000, to provide enough safety net for a person to quit their day job. Maybe higher if the opportunity cost of their day-job is higher)

If a thing is *really* important to you, you should consider the amount of effort that's going in, and be aware that this effort is a cost getting paid *somewhere* in the universe.

Maybe it's worth it to you to put in some portion of the herculean effort.

Maybe it's not - maybe the unfortunate equilibrium really is: "there was one person who was willing to put in 100 hours and a bunch of people who were willing to put in 1 or 2, but not enough to learn the skills necessary for those 100 hours to really work." It may be sad-but-true that it isn't actually worth it to any of the individuals involved to ensure the thing can continue running in it's original form.

But a thing to at least consider is whether, *long in advance* of the next organizers burning out, you should invest 20 or so hours gaining at least one of the skills that the organizers had developed. So that you can not just chip in with an hour or two of labor, but contribute one of the foundational building-blocks a given event, community or project needed to function.

# Against the Linear Utility Hypothesis and the Leverage Penalty

[Roughly the second half of this is a reply to: Pascal's Muggle]

There's an assumption that people often make when thinking about decision theory, which is that utility should be linear with respect to amount of stuff going on. To be clear, I don't mean linear with respect to amount of money/cookies/etc that you own; most people know better than that. The assumption I'm talking about is that the state of the rest of the universe (or multiverse) does not affect the marginal utility of there also being someone having certain experiences at some location in the uni-/multi-verse. For instance, if 1 util is the difference in utility between nothing existing, and there being a planet that has some humans and other animals living on it for a while before going extinct, then the difference in utility between nothing existing and there being n copies of that planet should be n utils. I'll call this the Linear Utility Hypothesis. It seems to me that, despite its popularity, the Linear Utility Hypothesis is poorly motivated, and a very poor fit to actual human preferences.

The Linear Utility Hypothesis gets implicitly assumed a lot in discussions of Pascal's mugging. For instance, in Pascal's Muggle, Eliezer Yudkowsky says he “[doesn't] see any way around” the conclusion that he must be assigning a probably at most on the order of $1/3\uparrow\uparrow\uparrow 3$ to the proposition that Pascal's mugger is telling the truth, given that the mugger claims to be influencing $3\uparrow\uparrow\uparrow 3$ lives and that he would refuse the mugger's demands. This implies that he doesn't see any way that influencing $3\uparrow\uparrow\uparrow 3$ lives could not have on the order of $3\uparrow\uparrow\uparrow 3$ times as much utility as influencing one life, which sounds like an invocation of the Linear Utility Hypothesis.

One argument for something kind of like the Linear Utility Hypothesis is that there may be a vast multiverse that you can influence only a small part of, and unless your utility function is weirdly nondifferentiable and you have very precise information about the state of the rest of the multiverse (or if your utility function depends primarily on things you personally control), then your utility function should be locally very close to linear. That is, if your utility function is a smooth function of how many people are experiencing what conditions, then the utility from influencing 1 life should be 1/n times the utility of having the same influence on n lives, because n is inevitably going to be small enough that a linear approximation to your utility function will be reasonably accurate, and even if your utility function isn't smooth, you don't know what the rest of the universe looks like, so you can't predict how the small changes you can make will interact with discontinuities in your utility function. This is a scaled-up version of a common argument that you should be willing to pay 10 times as much to save 20,000 birds as you would be willing to pay to save 2,000 birds. I am sympathetic to this argument, though not convinced of the premise that you can only influence a tiny portion of what is actually valuable to you. More importantly, this argument does not even attempt to establish that utility is globally linear, and counterintuitive consequences of the Linear Utility Hypothesis, such as Pascal's mugging, often involve situations that seem especially likely to violate the assumption that all choices you make have tiny consequences.

I have never seen anyone provide a defense of the Linear Utility Hypothesis itself (actually, I think I've been pointed to the VNM theorem for this, but I don't count that because it's a non-sequitor; the VNM theorem is just a reason to use a utility function

in the first place, and does not place any constraints on what that utility function might look like), so I don't know of any arguments for it available for me to refute, and I'll just go ahead and argue that it can't be right because actual human preferences violate it too dramatically. For instance, suppose you're given a choice between the following two options: 1: Humanity grows into a vast civilization of 10^100 people living long and happy lives, or 2: a 10% chance that humanity grows into a vast civilization of 10^102 people living long and happy lives, and a 90% chance of going extinct right now. I think almost everyone would pick option 1, and would think it crazy to take a reckless gamble like option 2. But the Linear Utility Hypothesis says that option 2 is much better. Most of the ways people respond to Pascal's mugger don't apply to this situation, since the probabilities and ratios of utilities involved here are not at all extreme.

There are smaller-scale counterexamples to the Linear Utility Hypothesis as well. Suppose you're offered the choice between: 1: continue to live a normal life, which lasts for n more years, or 2: live the next year of a normal life, but then instead of living a normal life after that, have all your memories from the past year removed, and experience that year again n more times (your memories getting reset each time). I expect pretty much everyone to take option 1, even if they expect the next year of their life to be better than the average of all future years of their life. If utility is just a naive sum of local utility, then there must be some year in which has at least as much utility in it as the average year, and just repeating that year every year would thus increase total utility. But humans care about the relationship that their experiences have with each other at different times, as well as what those experiences are.

Here's another thought experiment that seems like a reasonable empirical test of the Linear Utility Hypothesis: take some event that is familiar enough that we understand its expected utility reasonably well (for instance, the amount of money in your pocket changing by $5), and some ludicrously unlikely event (for instance, the event in which some random person is actually telling the truth when they claim, without evidence, to have magic powers allowing them to control the fates of arbitrarily large universes, and saying, without giving a reason, that the way they use this power is dependent on some seemingly unrelated action you can take), and see if you become willing to sacrifice the well-understood amount of utility in exchange for the tiny chance of a large impact when the large impact becomes big enough that the tiny chance of it would be more important if the Linear Utility Hypothesis were true. This thought experiment should sound very familiar. The result of this experiment is that basically everyone agrees that they shouldn't pay the mugger, not only at much higher stakes than the Linear Utility Hypothesis predicts should be sufficient, but even at arbitrarily large stakes. This result has even stronger consequences than that the Linear Utility Hypothesis is false, namely that utility is bounded. People have come up with all sorts of absurd explanations for why they wouldn't pay Pascal's mugger even though the Linear Utility Hypothesis is true about their preferences (I will address the least absurd of these explanations in a bit), but there is no better test for whether an agent's utility function is bounded than how it responds to Pascal's mugger. If you take the claim “My utility function is unbounded”, and taboo “utility function” and "unbounded", it becomes “Given outcomes A and B such that I prefer A over B, for any probability p>0, there is an outcome C such that I would take B rather than A if it lets me control whether C happens instead with probability p.” If you claim that one of these claims is true and the other is false, then you're just contradicting yourself, because that's what “utility function” means. That can be roughly translated into English as “I would do the equivalent of paying the mugger in Pascal's mugging-like situations”. So in Pascal's mugging-like situations, agents with unbounded utility functions don't look for clever

reasons not to do the equivalent of paying the mugger; they just pay up. The fact that this behavior is so counterintuitive is an indication that agents with unbounded utility functions are so alien that you have no idea how to empathize with them.

The “least absurd explanation” I referred to for why an agent satisfying the Linear Utility Hypothesis would reject Pascal's mugger, is, of course, the leverage penalty that Eliezer discusses in Pascal's Muggle. The argument is that any hypothesis in which there are n people, one of whom has a unique opportunity to affect all the others, must imply that a randomly selected one of those n people has only a 1/n chance of being the one who has influence. So if a hypothesis implies that you have a unique opportunity to affect n people's lives, then this fact is evidence against this hypothesis by a factor of 1:n. In particular, if Pascal's mugger tells you that you are in a unique position to affect 3↑↑↑3 lives, the fact that you are the one in this position is 1 : 3↑↑↑3 evidence against the hypothesis that Pascal's mugger is telling the truth. I have two criticisms of the leverage penalty: first, that it is not the actual reason that people reject Pascal's mugger, and second, that it is not a correct reason for an ideal rational agent to reject Pascal's mugger.

The leverage penalty can't be the actual reason people reject Pascal's mugger because people don't actually assign probability as low as 1/3↑↑↑3 to the proposition that Pascal's mugger is telling the truth. This can be demonstrated with thought experiments. Consider what happens when someone encounters overwhelming evidence that Pascal's mugger actually is telling the truth. The probability of the evidence being faked can't possibly be less than 1 in 10^10^26 or so (this upper bound was suggested by Eliezer in Pascal's Muggle), so an agent with a leverage prior will still be absolutely convinced that Pascal's mugger is lying. Eliezer suggests two reasons that an agent might pay Pascal's mugger anyway, given a sufficient amount of evidence: first, that once you update to a probability of something like 10^100 / 3↑↑↑3, and multiply by the stakes of 3↑↑↑3 lives, you get an expected utility of something like 10^100 lives, which is worth a lot more than $5, and second, that the agent might just give up on the idea of a leverage penalty and admit that there is a non-infinitesimal chance that Pascal's mugger may actually be telling the truth. Eliezer concludes, and I agree, that the first of these explanations is not a good one. I can actually demonstrate this with a thought experiment. Suppose that after showing you overwhelming evidence that they're telling the truth, Pascal's mugger says “Oh, and by the way, if I was telling the truth about the 3↑↑↑3 lives in your hands, then X is also true,” where X is some (a priori fairly unlikely) proposition that you later have the opportunity to bet on with a third party. Now, I'm sure you'd be appropriately cautious in light of the fact that you would be very confused about what's going on, so you wouldn't bet recklessly, but you probably would consider yourself to have some special information about X, and if offered good enough odds, you might see a good opportunity for profit with an acceptable risk, which would not have looked appealing before being told X by Pascal's mugger. If you were really as confident that Pascal's mugger was lying as the leverage prior would imply, then you wouldn't assume X was any more likely than you thought before for any purposes not involving astronomical stakes, since your reason for believing X is predicated on you having control over astronomical stakes, which is astronomically unlikely.

So after seeing the overwhelming evidence, you shouldn't have a leverage prior. And despite Eliezer's protests to the contrary, this does straightforwardly imply that you never had a leverage prior in the first place. Eliezer's excuse for using a leverage prior before but not after seeing observations that a leverage prior predicts are extremely unlikely is computational limitations. He compares this to the situation in which there is a theorem X that you aren't yet aware you can prove, and a lemma Y that you can

see is true and you can see implies X. If you're asked how likely X is to be true, you might say something like 50%, since you haven't thought of Y, and then when asked how likely X&Y is to be true, you see why X is probably true, and say something like 90%. This is not at all analogous to a “superupdate” in which you change priors because of unlikely observations, because in the case of assigning probabilities to mathematical claims, you only need to think about Y, whereas Eliezer is trying to claim that a superupdate can only happen when you actually observe that evidence, and just thinking hypothetically about such evidence isn't enough. A better analogy to the situation with the theorem and lemma would be when you initially say that there's a 1 in 3↑↑↑3 chance that Pascal's mugger was telling the truth, and then someone asks what you would think if Pascal's mugger tore a hole in the sky, showing another copy of the mugger next to a button, and repeating the claim that pushing the button would influence 3↑↑↑3 lives, and then you think “oh in that case I'd think it's possible the mugger's telling the truth; I'd still be pretty skeptical, so maybe I'd think there was about a 1 in 1000 chance that the mugger is telling the truth, and come to think of it, I guess the chance of me observing that evidence is around 10^-12, so I'm updating right now to a 10^-15 chance that the mugger is telling the truth.” Incidentally, if that did happen, then this agent would be very poorly calibrated, since if you assign a probability of 1 in 3↑↑↑3 to a proposition, you should assign a probability of at most 10^15 / 3↑↑↑3 to ever justifiably updating that probability to 10^-15. If you want a well-calibrated probability for an absurdly unlikely event, you should already be thinking about less unlikely ways that your model of the world could be wrong, instead of waiting for strong evidence that your model of the world actually is wrong, and plugging your ears and shouting “LA LA LA I CAN'T HEAR YOU!!!” when someone describes a thought experiment that suggests that the overwhelmingly most likely way the event could occur is for your model to be incorrect. But Eliezer perplexingly suggests ignoring the results of these thought experiments unless they actually occur in real life, and doesn't give a reason for this other than “computational limitations”, but, uh, if you've thought of a thought experiment and reasoned though its implications, then your computational limitations apparently aren't strict enough to prevent you from doing that. Eliezer suggests that the fact that probabilities must sum to 1 might force you to assign near-infinitesimal probabilities to certain easy-to-state propositions, but this is clearly false. Complexity priors sum to 1. Those aren't computable, but as long as we're talking about computational limitations, by Eliezer's own estimate, there are far less than 10^10^26 mutually disjoint hypotheses a human is physically capable of even considering, so the fact that probabilities sum to 1 cannot force you to assign a probability less than 1 in 10^10^26 to any of them (and you probably shouldn't; I suggest a “strong Cromwell's rule” that empirical hypotheses shouldn't be given probabilities less than 10^-10^26 or so). And for the sorts of hypotheses that are easy enough to describe that we actually do so in thought experiments, we're not going to get upper bounds anywhere near that tiny.

And if you do assign a probability of 1/3↑↑↑3 to some proposition, what is the empirical content of this claim? One possible answer is that this means that the odds at which you would be indifferent to betting on the proposition are 1 : 3↑↑↑3, if the bet is settled with some currency that your utility function is close to linear with respect to across such scales. But the existence of such a currency is under dispute, and the empirical content to the claim that such a currency exists is that you would make certain bets with it involving arbitrarily extreme odds, so this is a very circular way to empirically ground the claim that you assign a probability of 1/3↑↑↑3 to some proposition. So a good empirical grounding for this claim is going to have to be in terms of preferences between more familiar outcomes. And in terms of payoffs at familiar scales, I don't see anything else that the claim that you assign a probability of 1/3↑↑↑3 to a proposition could mean other than that you expect to continue to act as

if the probability of the proposition is 0, even conditional on any observations that don't give you a likelihood ratio on the order of 1/3↑↑↑3. If you claim that you would superupdate long before then, it's not clear to me what you could mean when you say that your current probability for the proposition is 1/3↑↑↑3.

There's another way to see that bounded utility functions, not leverage priors, are Eliezer's (and also pretty much everyone's) true rejection to paying Pascal's mugger, and that is the following quote from Pascal's Muggle: “I still feel a bit nervous about the idea that Pascal's Muggee, after the sky splits open, is handing over five dollars while claiming to assign probability on the order of 10^9/3↑↑↑3 that it's doing any good.” This is an admission that Eliezer's utility function is bounded (even though Eliezer does not admit that he is admitting this) because the rational agents whose utility functions are bounded are exactly (and tautologically) characterized by those for which there exists a probability p>0 such that the agent would not spend [fixed amount of utility] for probability p of doing any good, no matter what the good is. An agent satisfying the Linear Utility Hypothesis would spend $5 for a 10^9/3↑↑↑3 chance of saving 3↑↑↑3 lives. Admitting that it would do the wrong thing if it was in that situation, but claiming that that's okay because you have an elaborate argument that the agent can't be in that situation even though it can be in situations in which the probability is lower and can also be in situations in which the probability is higher, strikes me as an exceptionally flimsy argument that the Linear Utility Hypothesis is compatible with human values.

I also promised a reason that the leverage penalty argument is not a correct reason for rational agents (regardless of computational constraints) satisfying the Linear Utility Hypothesis to not pay Pascal's mugger. This is that in weird situations like this, you should be using updateless decision theory, and figure out which policy has the best a priori expected utility and implementing that policy, instead of trying to make sense of weird anthropic arguments before updatefully coming up with a strategy. Now consider the following hypothesis: “There are 3↑↑↑3 copies of you, and a Matrix Lord will approach one of them while disguised as an ordinary human, inform that copy about his powers and intentions without offering any solid evidence to support his claims, and then kill the rest of the copies iff this copy declines to pay him $5. None of the other copies will experience or hallucinate anything like this.” Of course, this hypothesis is extremely unlikely, but there is no assumption that some randomly selected copy coincidentally happens to be the one that the Matrix Lord approaches, and thus no way for a leverage penalty to force the probability of the hypothesis below 1/3↑↑↑3. This hypothesis and the Linear Utility Hypothesis suggest that having a policy of paying Pascal's mugger would have consequences 3↑↑↑3 times as important as not dying, which is worth well over $5 in expectation, since the probability of the hypothesis couldn't be as low as 1/3↑↑↑3. The fact that actually being approached by Pascal's mugger can be seen as overwhelming evidence against this hypothesis does nothing to change that.

Edit: I have written a follow-up to this.

# Thinking as the Crow Flies: Part 3 - Tokens, Syntax, and Expressions

# Preamble

In my last post, I discussed some basic logical connectives. Before moving on to more advanced topics, I want to discuss formal syntax. I think many people will find this post dry and boring, but I think it's neccessary to cover at some poin

Up till this point, I've been assuming an understanding which bears dwelling on. I've previously stated that a judgment is a series of mental tokens which one may declare. While I've *technically* adhered to that, it's obvious that I'm assuming more structure. Furthermore, I've made usage of variable binding and substitution, which have not been properly accounted for.

This short post is intended to specify the nature of syntax as used in this series. It will mostly be a recount of abstract syntax and binding according to the first chapter of *Practical Foundations for Programming Languages*, but with retrospect from the account of meaning explanations given earlier.

# Abstract Syntax Trees

Any given form of logic specifies various sorts of phrases which may be combined to form judgments. We'll distinguish between two different types of syntax, an informal "surface" syntax which is human readable, and a formal abstract syntax which is more well-defined. We'll be concerning ourselves with this abstract syntax in this post, treating the surface syntax as a mere readable short-hand for the abstract syntax.

In defining abstract syntax, we concern ourselves primarily with the rules by which we may combine phrases with each other. Since we already have an understanding of meaning explanation under our belt, we may use this to express the rules for correct syntax formation. At first, our phrases will be in the form of trees, called abstract syntax trees, or ASTs.

ASTs are made up of two main kinds of things, variables and operators. The variables may be placed at the leaves of an AST, while the operators make up the nodes. For the sake of restricting how syntax may be combined, we give every operator a sort. This sort will dictate the role of an operator by restricting how it may be combined with other operators.

Each operator comes equipped with a list of sorts, $s_1$, ... $s_n$, which tell us what sorts of ASTs can be plugged into the operator. In addition, each operator has an intrinsic sort s. When ASTs of the proper sorts are plugged into this operator, we obtain an AST of sort s. We denote the full arity $s(s_1, \ldots s_n)$.

Variables will be the most significant part of our discussion of abstract syntax. Briefly, a variable is a single token which is intended to be used as a marker for a place where a more concrete expression might be placed. As a result, expressions containing variables make up schemes of statements which may be specialized, via substitution, into more specific

statements. For example, in basic algebra, we may form polynomials such as $x^2 + 2x + 1$, which can be specialized by substitution of, say, 7 for x to obtain $7^2 + (2 \times 7) + 1$.

Abstract syntax trees are classified by sorts that divide ASTs into syntactic categories. As an example, the various true, prop, :, etc. operators will be placed into a different sort than $\lambda$, $\vee$, $\rightarrow$, etc. We would not have tried stating, for example, true true. Beyond not meaning anything, we will make this a syntactically incorrect statement.

Variables in abstract syntax trees range over sorts in the sense that only ASTs of the specified sort of the variable can be plugged into that variable. But the core idea carries over from school mathematics, namely that a variable is an unknown, or a place-holder, whose meaning is given by substitution.

As an example, consider a language of arithmetic expressions built from numbers, addition, and multiplication. The abstract syntax of such a language consists of a single sort Exp (standing for "expression") generated by the following:

- An operator 0 of arity N().
- An operator S of arity N(N).
- An operator num of arity Exp(N).
- Operators plus and times, both of arity $\mathrm{Exp}(\mathrm{Exp}, \mathrm{Exp})$.

As an example, the expression $2 + (3 \times x)$, which involves a variable, x, would be represented by the AST $\mathrm{plus}(\mathrm{num}(\mathrm{S}(\mathrm{S}(0))); \mathrm{times}(\mathrm{num}(\mathrm{S}(\mathrm{S}(\mathrm{S}(0)))); x))$ of sort Exp, under the assumption that x is also of sort Exp. For the sake of readability, all numbers will simply be written as numerals from now on. Because, say, num(4), is an AST of sort Exp, we may plug it in for x in the above AST to obtain$\mathrm{plus}(\mathrm{num}(2); \mathrm{times}(\mathrm{num}(3); \mathrm{num}(4)))$, which is written in informal, surface syntax as $2 + (3 \times 4)$. We may, of course, plug in any more complex ASTs of sort Exp for x instead.

We may give a meaning explanation for ASTs as follows;

> We may declare x to be an AST of sort s when we've declared x to be a variable of sort s .
>
> We may declare $\theta(a_1; a_2; \ldots a_n)$ to be an AST of sort s when we've declared $\theta$ to be an operator of arity $s(s_1, s_2, \ldots s_n)$ and have declared, for each i, $a_i$ to be an AST of sort $s_i$.

To fully explicate the system above, we have a meaning explanation like;

We may declare Exp to be a sort.

We may declare N to be a sort.

We may declare 0 an operator of arity N()

...

and this will fully specify our syntax.

Implicitly, for all the theories we've discussed so far, the exact abstract syntax is left implicit. But we may employ the following explanation;

We may declare Judgment to be a sort.

We may declare Proposition to be a sort.

We may declare hypothetical to be an operator of arity Judgment(Judgment, Judgment).

...

We may declare true an operator of arity Judgment(Proposition).

We may declare prop an operator of arity Judgment(Proposition).

We may declare ∧ an operator of arity Proposition(Proposition, Proposition)

We may declare ⊤ an operator of arity Proposition()

...

From these rules we can declare that

hypothetical ( prop ( ∧ ( A ; ⊤ ) ) ; true ( A ) )

is a judgment. Not one we can necessarily declare, but a well-formed judgment none the less. We typically write this as;

$$\frac{A \ \wedge \ \top \quad \text{prop}}{A \quad \text{true}}$$

which is our informal, surface syntax.

It's worth pointing out at this point that these sorts of things have a standard notation in the form of a BNF grammar. The above would typically be written more succinctly as;

Judgment, J ::=

true(P)

prop(P)

hypothetical(J; J)

...

Proposition, P ::=

⊤

∨ ( P ; P )

∧ ( P ; P )

...

we won't be spelling out grammars very much in this series, but, when need be, this syntax will be used. Consider it a short-hand for the kind of meaning explanation given before.

The meaning explanation for ASTs provides justification for a principle of reasoning called structural induction. Suppose that we wish to prove that some predicate holds of all ASTs of a given sort. To show this it is enough to consider all the ways in which it can be generated, and show that the predicate holds in each case under the assumption that it holds for its constituent ASTs (if any). So, in the case of the sort Exp just described, we must show that

- The predicate holds for any variable x of sort Exp.
- The predicate holds for any number, num(n).
- Assuming that the predicate holds for $a_1$ and $a_2$, prove that it holds for plus($a_1$; $a_2$) and times($a_1$; $a_2$).

Because these cases exhaust all formation possibilities, we are assured that the predicate holds for any AST of sort Exp. More generally, we'd need to prove that;

- $P_s(x)$ for any variable x of sort s.
- Given an operator θ of arity $s(s_1, \ldots s_n)$, assuming we have $Ps_1(a_1)$, ... $Ps_n(a_n)$, we must prove that $P_s(\theta(a_1; \ldots s_n))$.

This allows us to prove $P_s$ for all ASTs of sort s. We will use this technique to establish the well-behaviour of substitution over ASTs.

In general, when a grammar is defined we'll be reasoning about ASTs modulo some set of variables. Each AST has a set of variables appearing within it. We say that x is fresh in our AST if x does not appear in it, otherwise we say that x is free in our AST.

Given a sort s, a substitution a[b := c], where a is an AST of sort s, b is a variable of sort $s'$, and c is an AST of sort $s'$, is defined as;

$$x\,[\,x := b\,] \downarrow b$$

$x[y := b] \downarrow x$ when x and y are different variables

$$\theta(a_1; a_2; \ldots a_n)[\,x := c\,] \downarrow \theta(a_1[\,x := c\,]; a_2[\,x := c\,]; \ldots a_n[\,x := c\,])$$

By well-behaviour, I mean that, given a fixed AST, a substitution will always result in a unique new AST. To be more precise, we want to prove that

> Given an AST a with free variables $X, x$ and an AST b with free variables X, then there is a unique AST c for which $a[x := b] \downarrow c$.

To prove this, we consider the inductive cases for ASTs in general;

For the base case, we must consider what happens to a lone variable. That is, if we have a variable y which is free in $X, x$, then we have $y[x := b] \downarrow b$ in the case that y is x (in which case b is our unique c), and y otherwise (in which case y is our unique c).

For the inductive case, we must consider an operator of the form $\theta(a_1; a_2; \ldots a_n)$. Our inductive hypothesis is that there's a unique $c_i$ for each $a_i$ such that $a_i[x := b] \downarrow c_i$. In this case, we have

$$\theta(a_1; a_2; \ldots a_n)[\,x := b\,] \downarrow \theta(a_1[\,x := b\,]; a_2[\,x := b\,]; \ldots a_n[\,x := b\,])$$

by the definition of substitution, and

$$\theta(a_1[\,x := b\,]; a_2[\,x := b\,]; \ldots a_n[\,x := b\,]) \downarrow \theta(c_1; c_2; \ldots c_n)$$

by our inductive hypothesis. This means our unique c is $\theta(c_1; c_2; \ldots c_n)$.

That exhausts all formation possibilities. This establishes that substitution is well-behaved over ASTs, it will never get stuck.

# Abstract Binding Trees

Abstract binding trees, or ABTs, enrich ASTs with the ability to manage variables. The primary difference between an AST and an ABT is the ability to *bind* variables, allowing subtrees of an ABT to possess free variables which are not themselves free in the tree itself. The subtree where a bound variable is considered free is the *scope* of that variable. This is

useful for allowing an ABT to talk about indexed expressions and expression schemes. A primary consequence of this is the arbitrary nature of any particular choice of variable within a binding.

A common example of variable binding in ordinary mathematics is in integration. In the expression $\int f(x)dx$, we have an integral binding the variable x for use within the expression $f(x)$. The free variables of $f(x)$ are x, while there are no free variables in $\int f(x)dx$, assuming f is an operator. By the nature of this binding, the choice of x could have been otherwise without changing the expression. So we could rewrite the integral as $\int f(y)dy$. We would consider the two ABTs underlying these expressions to be identical. This identification is known as α-equivalence. Binding has a specific scope. For example, in the expression $x \int f(x)dx$, the first x is a variable which isn't bound by anything, while the second is the bound x of the integral. As just stated, this will be considered identical to $x \int f(y)dy$.

Furthermore, we may consider the expression $\int \int f(x, y)dxdy$ to be semantically different from $\int \int f(x, y)dydx$ as the order of bound variables is different.

Just like an AST, an ABT is made up of variables and operators, each having an associated arity. In the case of ABTs, it's also possible for an operator to bind variables of a particular sort within an argument. An operator of sort s with an argument of sort $s_1$ which binds variables of sort $b_1$, $b_2$, ... $b_n$ will have its arity denoted $s(b_{11}.\, b_{12}.\ldots.\, b_{1n}.\, s_1)$. Assuming that an integral binds something of sort Exp, we may denote its sort as Exp(Exp . Exp), and we may write the ABT of $\int f(x)dx$ as $\int(x.\, f(x))$. For the sake of readability, we will write

$$\vec{b_{11}}.\, b_{12}.\ldots.\, b_{1n}.\, s_1 \text{ as } \vec{b_1}.\, s_1.$$

Since we may rename variables within a binding without changing the ABT, we define a notion of variable replacement. Given a list of variables, $\vec{b}$, we may define a replacement of these variables as a list of pairs of variables, stating new and distinct names for the previous list. We denote such a replacement by ρ. This will be a function which, given a variable x in $\vec{b}$, will return a fresh variable ρ(x). We denote the application of such a replacement to an expression a with $\hat{\rho}(a)$.

Note that all variables within a renaming must be fresh relative to each-other to avoid binding conflicts. For example, we can't rename the variables in $\int \int f(x, y)dydx$ to

$\int \int f(z,z)dzdz$ since that changes the ABT, so$\rho \equiv \{x \leftrightarrow z, y \leftrightarrow z\}$ is not a valid renaming, but $\rho \equiv \{x \leftrightarrow z, y \leftrightarrow w\}$ is.

We may fully explain the meaning of ABTs as;

We may declare x to be an ABT of sort s when we've declared x to be a variable of sort s.

We may declare $\theta(b_1.a_1;\ldots b_n.a_n)$ to be an ABT of sort s when we've declared $\theta$ to be an operator of arity $s(b_{s1}.s_1,\ldots b_{sn}.s_n)$ and, for each i and all replacements $\rho_i$ for $b_i$ we have declared $\hat{\rho_i}(a_i)$ to be an ABT of sort $s_i$.

That renaming is significant. When we have an operator with sub-expressions that make use of bound variables, we don't want it to be a well-formed ABT under specific binding names, but rather, we want it to be well-formed under *all* binding names.

A version of structural induction called structural induction modulo fresh renaming holds for ABTs. The only difference is that we must establish our predicate for an ABT modulo renaming. To prove $P_s$ for all ABTs, we must prove that;

- $P_s(x)$ holds for any variable x of sort s
- Given an operator o of arity $s(b_{1s}.s_1,\ldots b_{ns}.s_n)$, assuming we have, for all i and all renaming $\rho_i$ of $b_i$ a proof of $P_{s_i}(\rho_i(a_i))$, we must prove that $P_s(\theta(b_1.a_1;\ldots b_n.s_n))$.

The second condition ensures that the inductive hypothesis holds for all fresh choices of bound variable names, and not just the ones actually given in the ABT.

We can now extend our grammars with rules that allow bound variables. Thus far, the most significant usages of variable binding are as part of generalized judgments and lambda expressions. Example grammar rules would look like;

We may declare $general_{prop}$ an operator of arity Judgment(Proposition . Judgment).

We may declare $general_{term}$ an operator of arity Judgment(Term . Judgment).

We may declare $\lambda$ an operator of arity Term(Term . Term).

Where Term is a sort in the grammar of type theory. Before now there was likely some confusion over exactly what general judgments were meant to bind. As it turns out, there are several judgments for each thing we may want to bind, all having the similar meaning explanations. Without a discussion of syntax, this ambiguity may have never been realized. Going forward, we will use operators which don't have the kind of over-loading that general judgments possess.

It will be useful to give a precise account of free occurrence. We can define a judgment, denoted $\mathrm{free}_x(a)$, indicating when a variable occurs freely within an ABT. That is, when a variable appears within an ABT, but isn't bound. We give the following meaning explanation;

> We may declare $\mathrm{free}_x(x)$.
>
> We may declare $\mathrm{free}_x(\theta(b_1.\, a_1; \ldots b_n.\, a_n))$ when, for some i, there is an $a_i$ such that, for each renaming $\rho_i$ of $b_i$, we may declare $\mathrm{free}_x(\hat{\rho_i}(a_i))$.

Some care is necessary to characterize substitution over ABTs. The most obvious issue is substitution of a variable who's name is already bound. Consider the case where x is bound at some point within a, then x does not occur free within a, and hence is unchanged by substitution. For example, $\int(x.\, a)[x := b] \downarrow \int(x.\, a)$, since there are no free occurrences of x in $x.\, a$, we can't declare $\mathrm{free}_x(\int(x.\, a))$. A different, perhaps more common, issue is free variable capture during substitution. For example, provided that x is distinct from y, $\int(y.\, \mathrm{plus}(x; y))[x := y]$ is not $\int(y.\, \mathrm{plus}(y; y))$, which confuses two different variables named y. Instead, it's simply undefined. It is, however, defined on a renaming of y. Capture can always be avoided by first renaming the bound variables. If we rename the bound variable y to $y'$ to obtain $\int(x'.\, \mathrm{plus}(x; y'))$, then $\int(x'.\, \mathrm{plus}(x; y'))[x := y]$ is defined, and reduces to $\int(x'.\, \mathrm{plus}(y; y'))$.

I've stated that ABTs which are α-equivalent should be considered as identical. The relation $a =_\alpha b$ of α-equivalence means that a and b are identical up to the choice of bound variable names. We may give α-equivalence the following meaning explanation;

> We may declare $x =_\alpha x$.

$\theta(\vec{b_1}.\vec{a_1}; \ldots \vec{b_n}.\vec{a_n}) =_\alpha \theta(\vec{b'_1}.\vec{a'_1}; \ldots \vec{b'_n}.\vec{a'_n})$ when we have declared, for every $i$ and every renaming $\rho_i$ of $b_i$ and renaming $\rho_i$ of $b_i$, that $\widehat{\rho_i}(a_i) =_\alpha \rho_i(a_i)$.

When we take this identification seriously, we can define substitution unproblematically;

$x\,[\,x := b\,] \downarrow b$

$x[y := b] \downarrow x$ when $x$ and $y$ are different variables

If we cannot, for any $b$ in any $\vec{b_i}$, declare $\text{free}_b(c)$, then

$\theta(\vec{b_1}.\vec{a_1}; \ldots \vec{b_n}.\vec{a_n})[x := c] \downarrow \theta(\vec{b_1}.\vec{a'_1}; \ldots \vec{b_n}.\vec{a'_n})$ where $a_i$ is simply $a_i$ if $x$ is in $\vec{b_i}$ and $a_i$ is $a_i[x := c]$ otherwise.

The conditions in that last rule prevent the problems discussed above. The condition on $b$ prevents naive variable capture, and can always be fulfilled by replacing all bound variables in our AST with fresh symbols. The conditions on the as prevent us from substituting variables which aren't free to be substituted.

Substitution is essentially performed on α-equivalence classes of ABTs rather than on an ABT itself. We can choose a capture-avoiding representative for a given class of α-equivalent ABTs and, any time we perform a substitution, we first canonicalize the binding names of the ABTs. This fixes all mentioned technical problems with substitution. There are several ways to do this in practice, for example by using de Bruijin indices.

# de Bruijin indices

If you're anything like me, much of the discussion of ABTs may not sit right with you. Something about considering ABTs up to swapping variables seems to be a bit of a hack, something which could easily lead to incorrect reasoning without one even taking notice. I mentioned that what should really be done is a process of canonicalization where a canonical representative of an ABT is chosen, and that de Bruijin indices could do this job. The point of this section is to describe how, exactly, this works.

A de Bruijin index is simply a numeral signifying a bound variable. 0 is the most recently bound, 1 the second most recent, 2 the third most recent, etc. If we wanted to, for sake of clarity, we could mark these numerals so that we don't confuse them with actual numbers. I'll use an over-line, $\overline{0}$, $\overline{1}$, $\overline{2}$, etc.

For a de Bruijin indexed ABT, we don't explicitly bind variables. Instead, for each place we look up the arity of the operator to know when variables are bound. For example, instead of writing $\lambda r.\, \mathrm{case}(r; e.\,(r, e); o.\,(r, o))$ we'd write $\lambda(\mathrm{case}(\bar{0}; (\bar{1}, \bar{0}); (\bar{1}, \bar{0})))$ notice that r becomes $\bar{0}$ outside of the subsequent bindings, and becomes $\bar{1}$ inside of them. After further bindings, the most recent variable becomes the second, third, etc. most recently bound variable.

Our notion of free variable has to be amended slightly. In the expression $\lambda((\bar{0}, \bar{1}))$, $\bar{1}$ is free while $\bar{0}$ is bound. We define the following, where $\mathrm{free}_n(x)$ judges that $\bar{n}$ is free in x;

We may declare $\mathrm{free}_n(\bar{n})$.

We may declare $\mathrm{free}_n(\theta(x_0; x_1; \ldots x_n))$ when we've declared $\theta$ to be of arity $s(b_1.\, s_1, b_2.\, s_2, \ldots b_n.\, s_n)$ and, for some i, we've declared $\mathrm{free}_{n+|b_i|}(x_i)$.

Where $|b_i|$ is the length of the list $b_i$. In the real world, $\mathrm{free}_n(x)$ would be implemented on a computer as a function that returns a boolean. As an example, we may declare $\mathrm{free}_0(\lambda((\bar{0}, \bar{1})))$ since we know that $\lambda$ has arity $\mathrm{Exp}(\mathrm{Exp}.\,\mathrm{Exp})$ and we may declare $\mathrm{free}_{0+1}((\bar{0}, \bar{1}))$. We may declare that since we know that , has arity $\mathrm{Exp}(\mathrm{Exp}; \mathrm{Exp})$, and we may declare (in this case for i = 1) $\mathrm{free}_1(\bar{1})$.

In the course of that validation, the variable we were testing the freedom of changed its denotation as we entered into variable bindings. This is where much of the confusion over de Bruijin indices arises.

Dealing with substitution is a rather subtle business, but once all the details are clarified, the algorithm is rather simple. Consider this naive evaluation of a substitution;

$$\lambda(\lambda(\bar{1})[\bar{0} := (\lambda(\bar{0}), \bar{0})])$$

$$\downarrow \lambda(\lambda(\bar{1}[\bar{0} := (\lambda(\bar{0}), \bar{0})]))$$

$$\downarrow \lambda(\lambda(\bar{1}))$$

Hmm... is that right? Let's consider what this is doing when the variables are given explicit names. So $\lambda(\lambda(\bar{1})[\bar{0} := (\lambda(\bar{0}), \bar{0})])$ should be $\lambda x.\,(\lambda y.\, x)[x := (\lambda z.\, z, x)]$. Well, clearly we did something wrong, our final evaluation shouldn't be $\lambda x.\, \lambda y.\, x$. We did not take into account

subsiquent bindings. When entering a binding, the variable we're substituting for goes from being the nth bound variable to the (n+1)th bound variable. Let's try again;

$$\lambda(\lambda(\overline{0})[\overline{0} := (\lambda(\overline{0}), \overline{0})])$$

$$\downarrow \lambda(\lambda(\overline{1}[\overline{1} := (\lambda(\overline{0}), \overline{0})]))$$

$$\downarrow \lambda(\lambda((\lambda(\overline{0}), \overline{0})))$$

Now, is that right? Our final evaluation was $\lambda x.\, \lambda y.\, (\lambda z.\, z, y)$. Well, that's not right. The second part of our pair started out being bound as x but ended up being bound as y. Clearly, we also need to increment the variables in what we're substituting as well whenever we enter into a binding.

$$\lambda(\lambda(\overline{0})[\overline{0} := (\lambda(\overline{0}), \overline{0})])$$

$$\downarrow \lambda(\lambda(\overline{1}[\overline{1} := (\lambda(\overline{1}), \overline{1})]))$$

$$\downarrow \lambda(\lambda((\lambda(\overline{1}), \overline{1})))$$

So, our final evaluation is now $\lambda x.\, \lambda y.\, (\lambda z.\, y, x)$. Wait, that's *still* not right. The first part of our pair had a variable bound as z, but this became a y after incrementing. This is because that variable wasn't free for the ABT $(\lambda(\overline{0}), \overline{0})$. So we see what we must do; only modify *free* variables during substitution.

$$\lambda(\lambda(\overline{0})[\overline{0} := (\lambda(\overline{0}), \overline{0})])$$

$$\downarrow \lambda(\lambda(\overline{1}[\overline{1} := (\lambda(\overline{0}), \overline{1})]))$$

$$\downarrow \lambda(\lambda((\lambda(\overline{0}), \overline{1})))$$

So our final evaluation is $\lambda x.\, \lambda y.\, (\lambda z.\, z, x)$, which is correct. The modification that we need to perform on our substituting expression is called *quotation*, and is a function defined like so;

$$\text{quote}_n^{\,i}(\overline{m}) \downarrow \overline{i + m} \text{ when } m \geq n.$$

$$\text{quote}_n^{\,i}(\overline{m}) \downarrow \overline{m} \text{ when } m < n.$$

$\text{quote}_n^i(\theta(x_0; \ldots x_m)) \downarrow \theta(\text{quote}^i_{n+|b_0|}(x_0); \ldots \text{quote}^i_{n+|b_m|}(x_m))$ when $\theta$ has arity $s(b_0.s_0; \ldots b_m.s_m)$.

Using this quotation function, we can define substitution as follows;

$\bar{x}[x := e] \downarrow e.$

$\bar{x}[y := e] \downarrow \bar{x}$ when $x \neq y$.

$\theta(x_0; \ldots x_m)[n := e] \downarrow \theta(x_0[n + |b_0| := \text{quote}^{|b_0|}_0(e)]; \ldots x_m[n + |b_m| := \text{quote}^{|b_m|}_0(e)])$ when $\theta$ has arity $s(b_0.s_0; \ldots b_m.s_m)$.

This works fine, but there's an extra hitch to working with de Bruijin indices, and that's dealing with the removal of variable bindings. Take the following reduction;

$$\text{ap}\ (\ \lambda\ x\ .\ y\ ;\ z\ )\ \downarrow\ y\ [\ x\ :=\ z\ ]$$

Notice that the x binding is being removed. That means that any subsequent binding in y will no longer be the nth binding, but will now be the (n-1)th binding. This means we need to actually decrement all the free variables in y by quoting with a -1.

$$\text{ap}\ (\ \lambda\ (\ y\ )\ ;\ z\ )\ \downarrow\ \text{quote}^{-1}_0\ (\ y\ [\ \overline{0}\ :=\ z\ ]\ )$$

Similarly, we have the meaning;

We may declare $|_x J$ when we may declare $J[x := E]$ for any expression E.

This must be modified when using de Bruijn indices. For example;

We may declare $|(\,)J)$ when we may declare $\text{quote}^{-1}_0\ (J[\bar{0} := E])$ for any expression E.

All this may seem rather complicated. A common quote goes "de Bruijn syntax is a Cylon detector". Despite that, de Bruijin indices are probably the simplest method for handling the canonicalization of ABTs, and this is common in actual computer implementations such as in Agda. That being said, we will not make mention of de Bruijin indices after this section. I think it's important to know about them, at least to have peace of mind. Using them allows

one to fix all possible ambiguities in variable bindings, and after a while they *do* become more intuitive, but they hardly assist readability.

# Writing Down Conversations

*Epistemic Status: Didn't think through exactly how I worded things.*

**tldr:** When you have insightful conversation, write it down and share it so people can build on it (instead of just sharing in person). Most of humanity's power comes from being able to build complex thoughts out of other thoughts and transmit them across the world.

This is a rehash/re-examining post. Related:

- Write Down Your Process (Zvi Mowshowitz)
- Some Thoughts on Public Discourse (Holden Karnofsky)
- Why and How to Name Things (Conor Moreton)
- Single Locus of Discussion (Anna Salamon)
- Return to Discussion (Sarah Constantin)
- [Edit: Turning Discussions into Blogposts by Brian Tomasik apparently covers this very topic. Still needed to be said again]

This is part 2 of N of my "Ray writes down conversations he had with people" series. It's also the most adorably meta of them.

A month ago I was talking with Oliver Habryka about why Less Wrong was important. (Call us biased if you will). One thing we both noted: in the days of yore, it seemed like a lot of prominent scholars/thinkers wrote down their insights and research on Less Wrong. Then, eventually they turned professional and joined official organizations whose job was to think fulltime.

*Also* over the past few years, those organizations (including but not limited to MIRI, CFAR, Givewell/OpenPhil) shifted from being younger-with-nothing-to-lose to older-with-reputations-to-safeguard, and their public facing tone seems to have shifted from "earnestly sharing thoughts as they come up" to "carefully crafted PR statements."

Thirdly, a lot of people moved to major geographic hubs, where it became easier to have in person conversations than to communicate via written blogpost. So... that's what people have tended to do.

I sympathize with the notion that people are busy and writing things up is a) time consuming and b) potentially risky. But I think the consequences of this are at least underweighted.

At *least*, I think people having informal conversations should make more of an effort to write those up in accessible form when there *aren't* reasons to be cautious.

Consequences of *not* writing stuff down include:

1. If you're not plugged into the personal-conversation-network, it's hard to keep up with a lot of collective insights relating to both rationality, effective altruism and x-risk. (This results in weird filtering effects which aren't the *worst* - being able to network your way is a credible signal of *something*. But I don't think it's the best possible filter)
2. It's actually fairly time consuming to propogate ideas via in-person conversation. Like, one of the *major* advantages of humanity is ideas being able to efficiently

spread via writing.
3. Another major advantage of writing is being able to *increase your collective working memory and build on ideas.* When much of your insights are developed via conversation, not only are you preventing far-away-people from building on your ideas, but you are hampering the ability of *yourself and your immediate colleagues* to build on those ideas. Writing things down (and then turning them into essays with titles that can be easily referenced) makes it easier to build complex models.

I think people feel a lot of pressure to write things up *well*, and as a result don't write things up at all. So my current take is, if you have a conversation that seems to contain important insights, err on the side of getting it written up *quickly* without as much regard for being timeless.

Later on if you think it deserves to be written up in a more timeless, scholarly form, you (or someone with more time) can still do that.

# TSR #5 The Nature of Operations

***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 # 5: The Nature of Operations

SUMMARY

# Religio → Strategy → Tactics → Operations

- Zhukov was pretty cool. You should def read to article to know why.
- Ops can be “hard to get” because they often seem too simple to be considered.
- Guidance

1. “Though operations cannot be understood by the individual pieces alone, the best practices of implementing and running the individual pieces are critical to learn.”
2. “Study and learn how different pieces of operations fit together over time, practice analyzing where one’s operations are the weakest, and how to improve operations across the spectrum.”

I really like using this hierarchy to think about things. Just remembering that these levels of planning exist helps me notice when I’m not acting on them. How does my religio inform my strategy? How does my strategy inform my tactics? How do my tactics inform my operations? When I’m falling apart, these stay as rhetorical questions. When I’m on top of things, they get answered.

Sebastian’s definition of Operations seems pretty accurate. *The coordination of tactics over time*. However, I prefer to think about operations via a question. ***What do you do to ensure that the things you intend to do actually get done?*** That seems to be at the core of operations, which seem to be at the core of getting good at things and living a quality life. Lots of low hanging fruit for improving one’s life is stuff you’ve heard before. Understanding operations lets you go, “Ohhhh, that’s why nothing worked. I was only trying to try, not actually doing.”

It seems like there are two big categories of operations. The first are operations that ensure you get a *particular thing* done. Examples would be a weightlifting regimen that a coach makes for you, a morning routine that you use to feel ready to tackle the day, or going grocery shopping every other friday to ensure you’ve got enough food. The second are operations that allow you to get a *particular class of things* done. Examples would be using a calendar to keep track of appointments, or using todoist to make sure miscellaneous errands get done.

I find it powerful just to be aware of what things I do and don’t have operations in place for. I’ve got solid operations to make sure my school work gets done during the hours I want it to get done. I’ve got okay ish operations to ensure that misc tasks get

done. I don’t have operations in place to eat healthy. I just recently realized that despite what I thought, I don’t have operations to ensure I make all my appointments (I put everything on my calendar, but it turns out my weeks are so routine that I don’t check it much besides when I’m planning the week).

This way of thinking let’s me start with a small bubble of control and slowly expand outwards.

So with all that in mind, here are a handful of questions to answer that might give you an idea of ways to become stronger.

1. **Do you have strong reasons to believe that you will execute the tactics most central to your current strategy?**
2. **What are the things you actually get done? What are the things that routinely fall through the cracks? How does the nature of your existing operations fail to support the things that aren’t happening?**
3. **How could you design new operations to increase the area of “Thing you can count on happening”?**

# Improvement Without Superstition

When you make continuous, incremental improvements to something, one of two things can happen. You can improve it a lot, or you can fall into superstition. I'm not talking about black cats or broken mirrors, but rather humans becoming addicted to whichever steps were last seen to work, instead of whichever steps produce their goal.

I've seen superstition develop first hand. It happened in one of the places you might least expect it – in a biochemistry lab. In the summer of 2015, I found myself trying to understand which mutants of a certain protein were more stable than the wildtype. Because science is perpetually underfunded, the computer that drove the equipment we were using was ancient and frequently crashed. Each crash wiped out an hour or two of painstaking, hurried labour and meant we had less time to use the instrument to collect actual data. We really wanted to avoid crashes! Therefore, over the course of that summer, we came up with about 12 different things to do before each experiment (in sequence) to prevent them from happening.

We were sure that 10 out of the 12 things were probably useless, we just didn't know which ten. There may have been no good reason that opening the instrument, closing, it, then opening it again to load our sample would prevent computer crashes, but as far as we could tell when we did that, the machine crashed far less. It was the same for the other eleven. More self-aware than I, the graduate student I worked with joked to me: "this is how superstitions get started" and I laughed along. Until I read two articles in The New Yorker.

In The Score (How Childbirth Went Industrial), Dr. Atul Gawande talks about the influence of the Apgar score on childbirth. Through a process of continuous competition and optimization, doctors have found out ways to increase the Apgar scores of infants in their first five minutes of life – and how to deal with difficult births in ways that maximize their Apgar scores. The result of this has been a shocking (six-fold) decrease in infant mortality. And all of this is despite the fact that according to Gawande, "[in] a ranking of medical specialties according to their use of hard evidence from randomized clinical trials, obstetrics came in last. Obstetricians did few randomized trials, and when they did they ignored the results."

Similarly, in The Bell Curve (What happens when patients find out how good their doctors really are), Gawande found that the differences between the best CF (cystic fibrosis) treatment centres and the rest turned out to hinge on how rigorously each centre followed the guidelines established by big clinical trials. That is to say, those that followed the accepted standard of care to the letter had much lower survival rates than those that hared off after any potentially lifesaving idea.

It seems that obstetricians and CF specialists were able to get incredible results without too much in the way of superstitions. Even things that look at first glance to be minor superstitions often turned out not to be. For example, when Gawande looked deeper into a series of studies that showed forceps were as good as or better than Caesarian sections, he was told by an experienced obstetrician (who was himself quite skilled with forceps) that these trials probably benefitted from serious selection effects (in general, only doctors particularly confident in their forceps skills volunteer for studies of them). If forceps were used on the same industrial scale as Caesarian sections, that doctor suspected that they'd end up worse.

But I don't want to give the impression that there's something about medicine as a field that allows doctors to make these sorts of improvements without superstition. In The Emperor of all Maladies, Dr. Siddhartha Mukherjee spends some time talking about the now discontinued practices of "super-radical" mastectomy and "radical" chemotherapy. In both treatments, doctors believed that if some amount of a treatment was good, more must be better. And for a while, it seemed better. Cancer survival rates improved after these procedures were introduced.

But randomized controlled trials showed that there was no benefit to those invasive, destructive procedures beyond that offered by their less-radical equivalents. Despite this evidence, surgeons and oncologists clung to these treatments with an almost religious zeal, long after they should have given up and abandoned them. Perhaps they couldn't bear to believe that they had needlessly poisoned or maimed their patients. Or perhaps the superstition was so strong that they felt they were courting doom by doing anything else.

The simplest way to avoid superstition is to wait for large scale trials. But from both Gawande articles, I get a sense that matches with anecdotal evidence from my own life and that of my friends. It's the sense that if you want to do something, anything, important – if you want to increase your productivity or manage your depression/anxiety, or keep CF patients alive – you're likely to do much better if you take the large scale empirical results and use them as a springboard (or ignore them entirely if they don't seem to work for you).

For people interested in nootropics, melatonin, or vitamins, there's self-blinding trials, which provide many of the benefits of larger trials without the wait. But for other interventions, it's very hard to effectively blind yourself. If you want to see if meditation improves your focus, for example, then you can't really hide the fact that you meditated on certain days from yourself [1].

When I think about how far from the established evidence I've gone to increase my productivity, I worry about the chance I could become superstitious.

For example, trigger-action plans (TAPs) have a lot of evidence behind them. They're also entirely useless to me (I think because I lack a visual imagination with which to prepare a trigger) and I haven't tried to make one in years. The Pomodoro method is widely used to increase productivity, but I find I work much better when I cut out the breaks entirely – or work through them and later take an equivalent amount of time off whenever I please. I use pomos only as a convenient, easy to Beemind measure of how long I worked on something.

I know modest epistemologies are supposed to be out of favour now, but I think it can be useful to pause, reflect, and wonder: when is one like the doctors saving CF patients and when is one like the doctors doing super-radical mastectomies? I've written at length about the productivity regime I've developed. How much of it is chaff?

It *is* undeniable that I am better at things. I've rigorously tracked the outputs on Beeminder and the graphs don't lie. Last year I averaged 20,000 words per month. This year, it's 30,000. When I started my blog more than a year ago, I thought I'd be happy if I could publish something once per month. This year, I've published 1.1 times per week.

But people get better over time. The uselessness of super-radical mastectomies was masked by other cancer treatments getting better. Survival rates went up, but when

the accounting was finished, none of that was to the credit of those surgeries.

And it's not just uselessness that I'm worried about, but also harm; it's possible that my habits have constrained my natural development, rather than promoting it. This *has* happened in the past, when poorly chosen metrics made me fall victim to Campbell's Law.

From the perspective of avoiding superstition: even if you believe that medicine cannot wait for placebo controlled trials to try new, potentially life-saving treatments, surely you must admit that placebo controlled trials are good for determining which things aren't worth it (take as an example the very common knee surgery, arthroscopic partial meniscectomy, which has repeatedly performed no better than sham surgery when subjected to controlled trials).

Scott Alexander recently wrote about an exciting new antidepressant failing in Stage I trials. When the drug was first announced, a few brave souls managed to synthesize some. When they tried it, they reported amazing results, results that we now know to have been placebo. Look. You aren't getting an experimental drug synthesized and trying it unless you're pretty familiar with nootropics. Is the state of self-experimentation really that poor among the nootropics community? Or is it really hard to figure out if something works on you or not [2]?

Still, reflection isn't the same thing as abandoning the inside view entirely. I've been thinking up heuristics since I read Dr. Gawande's articles; armed with these, I expect to have a reasonable shot at knowing when I'm at risk of becoming superstitious. They are:

- If you genuinely care only about the outcome, not the techniques you use to attain it, you're less likely to mislead yourself (beware the person with a favourite technique or a vested interest!).

- If the thing you're trying to improve doesn't tend to get better on its own and you're only trying one potentially successful intervention at a time, fewer of your interventions will turn out to be superstitions and you'll need to prune less often (much can be masked by a steady rate of change!).

- If you regularly abandon sunk costs ("You abandon a sunk cost. You didn't want to. It's crying."), superstitions do less damage, so you can afford to spend less mental effort on avoid them.

Finally, it might be that you don't care that some effects are placebo, so long as you get them and get them repeatedly. That's what happened with the experiment I worked on that summer. We knew we were superstitious, but we didn't care. We just needed enough data to publish. And eventually, we got it.

# Footnotes:

[1] Even so, there are things you can do here to get useful information. For example, you could get in the habit of collecting information on yourself for a month or so (like happiness, focus, etc.), then try several combinations of interventions you think might work (e.g. A, B, C, AB, BC, CA, ABC, then back to baseline) for a few weeks each. Assuming that at least one of the interventions doesn't work, you'll have a placebo to compare against. Although be sure to correct any results for multiple comparisons.

[2] That people still buy anything from HVMN (after they rebranded themselves in what might have been an attempt to avoid a study showing their product did no better than coffee) actually makes me suspect the latter explanation is true, but still.

# The map of "Levels of defence" in AI safety

One of the main principles of engineering safety is multilevel defence. When a nuclear bomb accidentally fell from the sky in the US, 3 of 4 defence levels failed. The last one prevented the nuclear explosion: https://en.wikipedia.org/wiki/1961_Goldsboro_B-52_crash

Multilevel defence is used a lot in the nuclear industry and includes different systems of passive and active safety, starting from the use of delayed neutrons for the reaction activation and up to control rods, containment building and exclusion zones.

Here, I present a look at the AI safety from the point of view of multilevel defence. This is mainly based on two of my yet unpublished articles: “Global and local solutions to AI safety” and “Catching treacherous turn: multilevel AI containment system”.

The special property of the multilevel defence, in the case of AI, is that the biggest defence comes from only the first level, which is AI alignment. Other levels have progressively smaller chances to provide any protection, as the power of self-improving AI will grow after it will break of each next level. So we may ignore all levels after AI alignment, but, oh Houston, we have a problem: based on the current speed of AI development, it seems that powerful and dangerous AI could appear within several years, but AI safety theory needs several decades to be created.

The map is intended to demonstrate a general classification principle of the defence levels in AI safety, but not to list all known ideas on the topic. I marked in “yellow” boxes, which are part of the plan of MIRI according to my understanding.

I also add my personal probability estimates as to whether each level will work (under the condition that AI risks are the only global risk, and previous levels have failed).

The principles of the construction of the map are similar to my “plan of x-risks prevention” map and my “immortality map”, which are also based around the idea of the multilevel defence.

pdf: https://goo.gl/XH3WgK

# Guarding Slack vs Substance

Builds on concepts from:

- Slack
- Goodheart's Imperius
- Nobody Does the Thing They are Supposedly Doing

**Summary:** If you're trying to preserve your sanity (or your employees') by scaling back on the number of things you're trying to do... make sure not to accidentally scale back on things that were important-but-harder-to-see, in favor of things that aren't as important but more easily evaluated.

*Epistemic Effort: Had a conversation, had some immediate instinctive reactions to it, did not especially reflect on it. Hope to flesh out how to manage these tradeoffs in the comments.*

---

Zvi introduced the term "Slack" to the rationaljargonsphere a few months ago, and I think it's the most *clearly useful* new piece of jargon we've seen in a while.

Normally, when someone coins a new term, I immediately find people shoehorning it into conversations and concepts where it doesn't quite fit. (I do this myself an embarrassing amount, and the underlying motivation is clearly "I want to sound smart" which bodes ill).

By contrast, I experienced an explosion of people jargon-dropping Slack into their conversations and *every single instance was valid*. Lack-of-slack was a problem loads of people had been dealing with, and having a handle for it was a perfect instance of a new name enabling higher level discussion.

This hints at something that should be alarming: "*slack*" *is a useful term because nobody has enough of it.*

In particular, it looks like many organizations I'm familiar with run at something like -10% slack, instead of the 40% slack that apparently is optimal across many domains.

Gworley noted in the comments of Zvi's post:

> If you work with distributed systems, by which I mean any system that must pass information between multiple, tightly integrated subsystems, there is a well understood concept of *maximum sustainable load* and we know that number to be roughly 60% of maximum possible load for all systems.
>
> The probability that one subsystem will have to wait on another increases exponentially with the total load on the system and the load level that maximizes throughput (total amount of work done by the system over some period of time) comes in just above 60%. If you do less work you are wasting capacity (in terms of throughput); if you do more work you will gum up the works and waste time waiting even if all the subsystems are always busy.
>
> We normally deal with this in engineering contexts, but as is so often the case this property will hold for basically anything that looks sufficiently like a distributed

> system. Thus the "operate at 60% capacity" rule of thumb will maximize throughput in lots of scenarios: assembly lines, service-oriented architecture software, coordinated work within any organization, an individual's work (since it is normally made up of many tasks that information must be passed between with the topology being spread out over time rather than space), and perhaps most surprisingly an individual's mind-body.
>
> "Slack" is a decent way of putting this, but we can be pretty precise and say you need ~40% slack to optimize throughput: more and you tip into being "lazy", less and you become "overworked".

I've talked with a few people people about burnout, and other ways that lack-of-slack causes problems. I had come to the conclusion that people should probably drastically-cut-back on the amount of things they're trying to do, so that they can afford to do them well, for the long term.

Oliver Habryka made a surprising case for why this might be a bad idea, at least if implemented carelessly. (The rest of this post is basically a summary of a month-ago conversation in which Oliver explained some ideas/arguments to me. I'm fairly confident I remember the key points, but if I missed something, apologies)

# Veneer vs Substance

## Example 1: Web Developer Mockups

*(This example is slightly contrived - a professional web developer probably wouldn't make this particular mistake, but hopefully illustrates the point)*

Say you're a novice web-developer, and a client hires you to make a website. The client doesn't understand anything about web development, but they can easily tell if a website is ugly or pretty. They give you some requirements, including 4 menu items at the top of the page.

You have a day before the first meeting, and you want to make a good first impression. You have enough time that you could build a site with a good underlying structure, but no CSS styling. You know from experience the client will be unimpressed.

So instead you throw together a quick-and-dirty-but-stylish website that meets their requirements. The four menu items flow beautifully across the top of the page.

They see it. They're happy. They add some more requirements to add more functionality, which you're happy to comply with.

Then eventually they say "okay, now we need to add a 5th menu-item."

And... it turns out adding the 5th menu item a) totally wrecks the visual flow of the page you designed, b) you can't even do it easily because you threw together something that manually specified individual menu items and corresponding pages, instead of an easily scalable menu-item/page system.

Your site looked good, but it wasn't actually built for the most important longterm goals of your client, and neither you nor your client noticed. And now you have *more* work to do than you normally would have.

# Example 2: Running a Conference

If you run a conference, people will *notice* if you screw up the logistics and people don't have food or all the volunteers are stressed out and screwing up.

They won't notice if the breaks between sessions are 10 minutes long instead of 30.

But much of the value of most conferences isn't the presentations. It's in the networking, the bouncing around of ideas. The difference between 10 minute breaks and 30 minute ones may be the difference between people actually being able to generate valuable new ideas together, and people mostly rushing from one presentation to another without time to connect.

"Well, simple", you might say. "It's not that hard to make the breaks 30 minutes long. Just do that, and then still put as much effort into logistics as you can."

But, would have you have *thought* to do that, if you were preoccupied with logistics?

How many other similar types of decisions are available for you to make? How many of them will you notice if you don't dedicate time to specifically thinking about how to optimize the conference for producing connections, novel insights and new projects?

Say your default plan is to spend 12 hours a day for three months working your ass off to get *both* the logistics done and to think creatively about what the most important goals of the conference are and how to achieve them.

(realistically, this probably isn't actually your default plan, because thinking creatively and agentily is pretty hard and people default to doing "obvious" things like getting high-profile speakers)

But, you've also read a bunch of stuff about slack and noticed yourself being stressed out a lot. You try to get help to outsource one of the tasks, but getting people you can really count on is hard.

It looks like you need to do a lot of the key tasks yourself. You're stretched thin. You've got a lot of people pinging you with questions so you're running on manager-schedule instead of setting aside time for deep work.

You've run conferences before, and you have a lot of visceral experiences of people yelling at you for not making sure there was enough food and other logistical screwups. You have a lot of *current* fires going on and people who are yelling about them *right now.*

You *don't* have a lot of salient examples of people yelling at you for not having long enough breaks, and nobody is yelling at you *right now* to allocate deep work towards creatively optimizing the conference.

So as you try to regain sanity and some sense of measured control, the things that tend to get dropped are the things *least visible,* without regard for whether they are the most substantive, valuable things you could have done.

# So What To Do?

Now, I *do* have a strong impression that a lot of organizations and people I know are running at -10% slack. This is for understandable reasons: The World is On Fire, metaphorically and literally. There's a long list of Really Important Things that need doing.

Getting them set in motion *soon* is legitimately important.

A young organization has a *lot* of things they need to get going at once in order to prove themselves (both to funders, and to the people involved).

There aren't too many people who are able/willing to help. There's even fewer people who demonstratably can be counted on to tackle complex tasks in a proactive, agenty fashion. Those people end up excessively relied upon, often pressured into taking on more than they can handle (or barely exactly how much they can handle and then as soon as things go wrong, other failures start to snowball)

*[note: this is not commentary on any particular organization, just a general sense I get from talking to a few people, both in the rationalsphere but also generally in most small organizations]*

What do we do about this?

Answers would probably vary based on specific context. Some obvious-if-hard answers are obvious-if-hard:

- Try to buy off-the-shelf solutions for things that off-the-shelf-solutions exist for. (This runs into a *different* problem which is that you risk overpaying for enterprise software that isn't very good, which is a whole separate blogpost)
- Where possible, develop systems that dramatically simplify problems.
- Where possible, get more help, while generally developing your capacity to distribute tasks effectively over larger numbers of people.
- Understand that if you're pushing yourself (or your employees) for a major product release, you're not actually *gaining* time or energy - you're borrowing time/energy from the future. If you're spending a month in crunch time, expect to have a followup month where everyone is kinda brain dead. This may be worth it, and being able to think more explicitly about the tradeoffs being made may be helpful.

You're probably doing things like that, as best you can. My remaining thought is something like "do *fewer* things, but give yourself a lot more time to do them." (For example, I ran the 2012 Solstice almost entirely by myself, but I gave myself an entire year to do it, and it was my only major creative project that year)

If you're a small organization with a lot of big ideas that all feel interdependent, and if you notice that your staff are constantly overworked and burning out, it may be necessary to prune those ideas back and focus on 1-3 major projects each year that you have can afford to do well *without* resorting to crunch time.

Allocate time months in advance (both for thinking through the creative, deep underlying principles behind your project, as well as setting logistical systems in motion that'll make things easier).

None of this feels like a *satisfying* solution to me, but all feels like useful pieces of the puzzle to have in mind.

# Success and Fail Rates of Monthly Policies

*Inspired by Updates from Boston.*

One of the more valuable interventions I tested in the last couple years is the idea of setting a "Monthly Policy" and a "Monthly Theme." I'll share the basic process briefly, and then share a couple counterintuitive lessons with implications for self-improvement.

**Monthly Theming**

I've gotten immense value out of naming my months an uncommon phrase, and trying to live in accordance with that theme for the entirety of one month. For reference, here's my monthly themes for each month of 2017 —

1. Structured Data
2. Do Less, Better
3. Sentry Ops II
4. Repertoire
5. Excellence
6. Impact Floor
7. Glorious Neutrality
8. Acts
9. Boring Methodical Execution
10. Make List, Run List
11. Trust the Process
12. Build Supply

So, in January, "Structured Data," I focused on building out a more complete dataset that I could run for the whole year — setting up systems for tracking every minute, tracking how long projects take to complete, upgrading personal finance, tracking fitness and macronutrient consumption, and a few things like that.

In February, "Do Less Better," I focused on having less projects open at the same time, doing less activities overall, and more carefully choosing activities for upside/gains while also focusing on spending more time to increase output quality of all the work I did choose to do.

Last month, November, was "Trust the Process" — October had been a month with lots of travel that had heavily disrupted my core habits, sleep schedules, etc. In November, I focused on doing nothing particularly expansive and instead focused on resetting and re-installing every best practice I know.

This month, I'm focused on both stockpiling all common physical supplies and improving procurement (food, household supplies, medicine, etc) as well as increasing my backlog of intellectual work (essays written and edited that are in backlog ready to publish, all the logistical work for events pre-done with reminder and confirmation messages scheduled, etc).

Monthly Theming is great, because it's almost free. I think you could set a theme without much reflection in just 15-30 minutes and get some benefit, but doing a full

review takes me only 3-4 hours to do rigorously. I wind up making more correct decisions throughout the month, and some of the policies stick with me forever.

For instance, "Impact Floor" was about aiming to ensure I got 5 excellent hours in every single day and designing requisite plans and changing my workflows around that. The 3-4 hour investment to analyze and realize that this was important paid off in many more good hours that month, and I kept roughly half of the policies I experimented with and kept running them every month going forwards.

**How I Set Monthly Policies and Themes**

The first thing I do is a **Monthly Review**. This is very factual and not particularly subjective. Most of it requires very little judgment.

I first review last month's policy, and write down which aspects of it I adhered to or did not.

I then look at all the projects I did the previous month, and I literally look at every single day to identify what I did on each of them the past month. I look at all the data I keep and glance over it, looking for anything unusual.

Then I write four things —

1. Problem/Opportunity/Hypothesis
2. Short Policy Statement
3. Tools and Operationalization
4. Behavior Change

The first is where I write whatever I see as the current largest unsolved problem, the current largest opportunity, and/or any hypotheses I have about things I could or ought to be doing.

Then I write a policy statement. This takes the longest to think through and weigh what I want to work on. Some policy statements are very straightforward — "Build Supply" for this month is very straightforward — whereas others are more abstract and take a while to nail down. For instance, "Glorious Neutrality" was about repeatedly looking to reach neutrality and equanimity from any place of negative affect, but in a calm and expansive sort of way. The key line from that entry:

> *[Glorious Neutrality] means rapid recovery, rapid recentering, and vigilance to turn fight/flee/thrash reactions into productive time or leisure time.*

Then I write down what external factors and tools I'll need to develop or install into my life for that particular month, as well as what behavior changes I'll test out for that month.

To use Glorious Neutrality as an example again, I picked a number of interesting books and audiobooks at the start of the month, got a coworking membership for the month, and set up a series of work meetings / jam sessions with a couple productive friends.

This has been a very worthwhile usage of 3-4 hours per month.

**Success and Failure Rates of Monthly Themes**

I've now been running something resembling monthly policies since February 2016; this is the 23rd month I've done something like that. It was very much a work-in-progress and didn't hit its current form until January of this year.

Looking at my monthly policies from this year, the results are —

Large Successes: Structured Data (Janaury), Do Less Better (February), Impact Floor (June), Acts (August), Boring Methodical Execution (September), Trust the Process (November)

Failures: Sentry Ops II (March), Make List Run List (October)

Mixed: Repertoire (April), Excellence (May), Glorious Neutrality (July)

Of the 11 completed months of this year, that's 7 successes, 2 outright failures, and 2 middling underperforming failure-ish things.

**If you're counting then, 64% of monthly policies produce a large success.**

On average, I test between 5 and 12 implementation items each month between tools and behavior change. Here's November's ("Trust the Process") —

> *Tools and Operationalization: Get back on -I and -II Lights; Create a "Daily Calibration Template" for my journal; On Calibration, include tuning lights and develop repertoire; Get back on Project Speed of Execution tracking; Do rounds of formal Cycles again; Maybe put some jams with other people on the calendar.*
>
> *Behavior Change: I'll need to include countermeasure-planning in Daily Calibration; That should include leisure and fallback plans; Persistence is in order, especially around getting Cycles done; Spend a lot of time studying my behavior and ensure I adhere to it.*

Interestingly, **I only keep doing 1-2 items on average from each month.**

That means that, on average, 60%-90% of items I experiment with are discarded after being tested for 30 days.

**Implications**

We know that, due to loss aversion and a variety of cognitive biases, people tend to experience failure as more painful and distressing than success feels good and uplifting.

Likewise, it stands to reason that a month that was a failure is also likely a time period where someone has lower mental and physical resources and wellbeing — a month where you get sick is more likely to be a failed month while simultaneously reducing capacity/capability for next month.

I suspect that these types of self-improvement initiatives are likely to be abandoned after a run of 1-2 bad months, despite being useful overall. For a person to stick with monthly planning and theming, they'd need to have either an unusually good run to start it, or pre-committing to sticking with it during failed months.

Most importantly, **the realization that 60%-90% of behavior interventions aren't worth keeping was insightful and has a lot of implications.** Most people,

I suspect, probably only install one or a small handful of behavior changes and tools at a time.

But from my experiments, **the average individual tool or experimental new behavior *is not worth running going forwards.***

If this held true for other people, then it means that someone who attempts to install a single new tool or technology, or uptake a single new habit or behavior change, is far more likely than not doing something that isn't worth the cost of maintaining going forwards.

This could very easily lead to a person becoming discouraged by self-improvement experiments. I didn't intentionally set out to test 5-12 interventions every single month and keep the best of them — rather, they flowed from experiments along a common theme in order to try to actualize that theme.

For someone trying only a single behavior at a time, it's likely that *it would make perfect sense to abandon that behavior* — but, if they're only trying behaviors or tools one at a time, it's actually likely that they hit a run of failures on their first few attempts.

This would seem to be demoralizing, and discourage experimentation going forwards.

**Tentative Conclusions**

I find setting a "Monthly Theme" — basically, naming any given month a unique name that ties into a set of behavior changes and attentional focus — to be a very inexpensive way to get some gains. Try it out, if you like.

I do a rigorous review of the past month, and list out a mix of potential problems/opportunities/hypotheses before choosing a policy and theme for the next month. I've found the 3-4 hours that this takes to do, once per month, to be a very good use of time.

Explicitly setting out "tools and operationalization" and "behavior change" to support the theme and policy has been useful.

My experience has been that the majority of tools tested and experimental behavior changes are not worth keeping going forwards, yet, since I discover 1-2 worthwhile things to keep each month, the majority of months feel like a "win."

This might suggest that trying a "basket of changes around a single theme" to be a more reliable way to identify 1-2 good changes to keep going forwards, whereas trying to make individual piecemeal changes might result in seeming failure and demoralization.

I plan to keep running monthly themes and policies for the foreseeable future; it's been good for me. If you try it or have tried similar things, I'd love to hear about it. Questions and feedback are very welcome.

# Towards a Rigorous Model of Virtue-Signalling

The concept of "virtue signalling" has been a bit polarizing. Some people find it seems to explain a lot of the world; others find its lack of precision makes it almost an empty insult against others' behavior.

The model I'm about to propose doesn't capture everything people mean by "virtue signalling"; I'm not sure it even captures the heart of it. But it does demonstrate a rigorous basis for something that deserves the name "virtue signalling".

Suppose a pool of 100 players are assigned to play iterated Prisoner's Dilemma-like games, with the following properties:

0) We will relabel 'cooperate' and 'defect' actions as 'unselfish' and 'selfish', to better reflect the varying payoff matrices.

1) A match consists of 100 games.

2) 99 of the games are Deadlock, in which both players playing selfish is optimal for both players. Payoff matrix [[(1, 1), (3, 0)], [(0, 3), (2, 2)]]

3) One game, sprinkled in at random, is a Prisoner's Dilemma, with very high-magnitude payoffs, and in particular very negative ones if the other player plays selfish. Payoff matrix [[(0, 0)(100, -1500)], [(-1500, 100), (-1400, -1400)]

4) Playing a match with someone who will play selfish on a Prisoner's Dilemma is negative expected value, regardless of your play or their play on Deadlocks.

4) Playing a match with someone who plays unselfish on PD is positive expected value if you both play unselfish on Deadlocks, but of course even more positive if you both play selfish on Deadlocks.

Now suppose that the players face one another in a Round Robin structure, but get to accept or decline each match after the first. To inform their decision, they can see limited information on the other player: specifically, whether or not they have ever made a selfish play.

Now let us consider this tournament as a measure of a group's coordination ability. The least coordinated possible group would play straight selfish on their first match, then refuse to play any more matches, coming out with a total of (99 Deadlocks x 2 points per Deadlock - 1400 points per Prisoner's Dilemma x 1 Prisoner's Dilemma) = -1202 points per member. An optimally coordinated group, in contrast, would play selfish on every deadlock and unselfish on every Prisoner's Dilemma and play every match; each member would net (99 matches) x (99 Deadlocks x 2 points per deadlock + 1 PD x 0 points per PD) = 19602 points.

But the above strategy is vulnerable to defectors. Attempting the above strategy in a population with X defectors would give an honest player ((99 - X) matches with nice players) x (99 x 2)) + (X matches with defectors x -1302) = (99 - X) x 198 - 1302 X = 19602 - 1500X, while a defector gets (99 - (X-1)) matches with nice players x (99 x 2 + 100) + (X - 1) matches with defectors x -1202 = (100 - X) x 298 + (X - 1) x -1202 =

31002 - 1500X, which is strictly superior. So the optimal strategy is not a Nash equilibrium.

So let's consider an intermediate level of coordination, one willing to extend trust provisionally. 'Good' players will always play unselfish and decline games with those who have ever played selfish. If all players are 'good', they will each get (99 matches x 99 Deadlocks per match x 1 point per Deadlock) = 9801 points. A would-be defector will only get 397 points, so there's no incentive to defect. This is, I suspect but cannot prove, the optimal Nash equilibrium for the broader game; and all players get a noticeably suboptimal return in order to prove that they will not defect. That, I say, is virtue signalling.

# Pascal's Muggle Pays

Reply To (Eliezer Yudkowsky): Pascal's Muggle Infinitesimal Priors and Strong Evidence

Inspired to Finally Write This By (Lesser Wrong): Against the Linear Utility Hypothesis and the Leverage Penalty.

The problem of Pascal's Muggle begins:

> Suppose a poorly-dressed street person asks you for five dollars in exchange for doing a googolplex's worth of good using his Matrix Lord powers.

"Well," you reply, "I think it very improbable that I would be able to affect so many people through my own, personal actions – who am I to have such a great impact upon events? Indeed, I think the probability is somewhere around one over googolplex, maybe a bit less. So no, I won't pay five dollars – it is unthinkably improbable that I could do so much good!"

"I see," says the Mugger.

At this point, I note two things. I am not paying. And my the probability the mugger is a Matrix Lord is *much higher* than five in a googolplex.

That looks like a contradiction. It's positive expectation to pay, by *a lot,* and I'm not paying.

Let's continue the original story.

> A wind begins to blow about the alley, whipping the Mugger's loose clothes about him as they shift from ill-fitting shirt and jeans into robes of infinite blackness, within whose depths tiny galaxies and stranger things seem to twinkle. In the sky above, a gap edged by blue fire opens with a horrendous tearing sound – you can hear people on the nearby street yelling in sudden shock and terror, implying that they can see it too – and displays the image of the Mugger himself, wearing the same robes that now adorn his body, seated before a keyboard and a monitor.

"That's not actually me," the Mugger says, "just a conceptual representation, but I don't want to drive you insane. Now give me those five dollars, and I'll save a googolplex lives, just as promised. It's easy enough for me, given the computing power my home universe offers. As for why I'm doing this, there's an ancient debate in philosophy among my people – something about how we ought to sum our expected utilities – and I mean to use the video of this event to make a point at the next decision theory conference I attend. Now will you give me the five dollars, or not?"

"Mm… no," you reply.

"*No?*" says the Mugger. "I understood earlier when you didn't want to give a random street person five dollars based on a wild story with no evidence behind it. But now I've offered you evidence."

"Unfortunately, you haven't offered me *enough* evidence," you explain.

I'm paying.

So are you.

What changed?

I

The probability of Matrix Lord went up, but the odds were already there, and he's probably not a Matrix Lord (I'm probably dreaming or hypnotized or nuts or something).

At first the mugger *could benefit by lying to you.* More importantly, people other than the mugger could benefit by trying to mug you and others who reason like you, if you pay such muggers. They can exploit taking large claims seriously.

Now the mugger *cannot benefit by lying to you.* Matrix Lord or not, there's a cost to doing what he just did and it's higher than five bucks. He can extract as many dollars as he wants in any number of ways. A decision function that pays the mugger need not create opportunity for others.

I pay.

In theory Matrix Lord could derive some benefit like having data at the decision theory conference, or a bet with another Matrix Lord, and be lying. Sure. But if I'm even 99.999999999% confident this isn't for real, that seems nuts.

(Also, he could have gone for way more than five bucks. I pay.)

(Also, this guy gave me way more than five dollars worth of entertainment. I pay.)

(Also, this guy gave me way more than five dollars worth of good story. I pay.)

II

The leverage penalty is a crude hack. Our utility function is given, so our probability function had to move or the Shut Up and Multiply would do crazy things like pay muggers.

The way out is our decision algorithm. As per Logical Decision Theory, our decision algorithm is correlated to lots of things, including the probability of muggers approaching you on the street and what benefits they offer. The reason real muggers use a gun rather than a banana is mostly that you're far less likely to hand cash over to someone holding a banana. The fact that we pay muggers holding guns is why muggers hold guns. If we paid muggers holding bananas, muggers would happily point bananas.

There is a natural tendency to slip out of Functional Decision Theory into Causal Decision Theory. If I give this guy five dollars, how often will it save all these lives? If I give five dollars to this charity, what will that marginal dollar be spent on?

There's a tendency for some, often economists or philosophers, to go all lawful stupid about expected utility and *berate* us for *not* making this slip. They yell at us for voting, and/or asking us to justify not living in a van down by the river on microwaved ramen noodles in terms of our expected additional future earnings from our resulting increased motivation and the networking effects of increased social status.

To them, we must reply: We are choosing the logical output of our decision function, which changes the probability that we're voting on reasonable candidates, changes the probability there will be mysterious funding shortfalls with concrete actions that won't otherwise get taken, changes the probability of attempted armed robbery by banana, and changes the probability of random people in the street claiming to be Matrix Lords. It also changes lots of other things that may or may not seem related to the current decision.

Eliezer points out humans have bounded computing power, which does weird things to one's probabilities, especially for things that can't happen. Agreed, but you can defend yourself without making sure you never consider benefits multiplied by $3\uparrow\uparrow\uparrow 3$ without also dividing by $3\uparrow\uparrow\uparrow 3$. You can *have a logical algorithm* that says *not to treat differently* claims of $3\uparrow\uparrow\uparrow 3$ and $3\uparrow\uparrow\uparrow\uparrow 3$ if the justification for that number is someone telling you about it. Not because the first claim is so much less improbable, but because you don't want to get hacked in this way. That's way more important than the chance of meeting a Matrix Lord.

Betting on your beliefs is a great way to improve and clarify your beliefs, but you must think like a trader. There's a reason logical induction relies on markets. If you book bets on your beliefs at your fair odds without updating, you will get dutch booked. Your decision algorithm should not accept all such bets!

People are *hard* to dutch book.

Status quo bias can be thought of as *evolution's solution to not getting dutch booked.*

III

Split the leverage penalty into two parts.

The first is 'don't reward saying larger numbers'. Where are these numbers coming from? If the numbers come from math we can check, and we're offered the chance to save 20,000 birds, we can care much more than about 2,000 birds. A guy designing pamphlets picking arbitrary numbers, not so much.

Scope insensitivity can be thought of as *evolution's solution to not getting Pascal's mugged.* The one child is real. Ten thousand might not be. Both scope insensitivity and probabilistic scope sensitivity get you dutch booked.

Scope insensitivity and status quo bias cause big mistakes. We must fight them, but by doing so we make ourselves vulnerable.

You also have to worry about fooling *yourself.* You don't want to give *your own brain* reason to cook the books. There's an elephant in there. If you give it reason to, it can write down larger exponents.

The second part is applying Bayes' Rule properly. Likelihood ratios for seeming high leverage are usually large. Discount accordingly. How much is a hard problem. I won't go into detail here, except to say that if calculating a bigger impact doesn't increase how excited you are about an opportunity, *you are doing it wrong.*

# Can we see light?

Is "visible" light, actually visible? Claiming that visible light is called visible light and therefore it must be visible, is circular reasoning. This question is not about the definition of visible, because in that regard, light shows none of the characteristics of visible objects. Light is in fact, what makes objects visible.

Now I'm not talking about wavelengths we can't detect or even light that doesn't strike our eyes. I'm specifically referring to detectable light that strikes our retina. Many will see this as a futile argument about definitions until they actually grasp the differences and realize the implications.

The purpose of vision, what gives us an evolutionary advantage, is that it allows us to see things. For those unfamiliar with the concept of indirect realism, here's a link. http://cns-alumni.bu.edu/~slehar/webstuff/book/chap1.html . What we see are objects, like predators, food and possible mates. These objects exist in objective reality (outside our heads) but we perceive our brains representation of these objects in our subjective reality (the reality we perceive inside our heads).

DETECTION IS NOT PERCEPTION.

Our eyes detect light, but detection is a mechanical process of which we are not directly conscious of. We can deduce that the process is occurring due to the fact that we see objects and understand the visual process. Seeing something, on the other hand, is a conscious process. We sometimes don't even see things which are right in front of our eyes and I don't mean figuratively. Have you every moved something out of the way while looking in the fridge, when the thing you're looking for is the thing you moved? Seeing is not perceiving. Perception is consciously seeing something. Detection is not perception.

DO WE SEE OBJECTS OR LIGHT.

The most common held belief is that we perceive light, not objects. As far as detection goes, that's true. Our eyes detect light that strikes the retina. This begins the physiological aspect of vision. Phototransduction, electrochemical impulses travelling to the visual cortex via the optic nerve, the subconscious creating visual representations and sensations. All these processes are subconscious. It's only then that conscious perception comes in. And what we consciously perceive are the objects, from which the detected light, originates. So we do not see light, we see the (brains representation) objects.

LIGHT, LIGHT AND LIGHT.

Light, the word, has many meanings and this adds to the confusion of whether we see light or not. We have heavy and light, darkness and light, electromagnetic radiation and figuratively, seeing the light, which represents comprehension. We can see if something is heavy or light. We can see brightness. But brightness is not a property of electromagnetic radiation. Brightness is a sensation, like colors. If colors are the interpretation of lights wavelength, then brightness is the brains interpretation of lights amplitude. Now many people believe seeing light (brightness) is seeing light. That if you shine a laser into your eye, you're seeing light. In a sense, they're right. We are perceiving brightness. But this brightness is a result of our cones being -saturated. If you look at a 60w globe and hold your gaze steady, your eyes will adjust to the

brightness. After a few seconds you begin to see the element from which the light emanates. Brightness is an obstruction, preventing us from seeing the object itself. Brightness is phenomenal in nature. Not a property of light itself.

SEE AND SEE.

The word see, too has different meanings. The expression "to see the future " has two interpretations. One as a psychic, having visions of the future, and one as a visionary, able to predict future trends. To consciously perceive or to consciously conceive. So we can see "conceive " light but not perceive light.

WHAT DO WE ACTUALLY DETECT?

A photon is a boson particle. It can't actually be detected. What we detect is when a photon strikes something. It's the collision we detect. At the moment of collision, a photon no longer exists. Everything we know about light, is deduced by detecting these collisions. Light itself is undetectable. It neither emits, nor reflects anything which would allow us to detect it.

# A List Of Questions & Exercises For Reviewing Your Year

As the year comes to an end, I thought it might be useful to share some questions that might help with reviewing your past year.

This is an adapted post, based on what I have written here.

And, I've put together an interactive suite so you can answer the questions in a more frictionless way. It's on PH right now. I hope the exercise increases your utilotons.

Before starting, here are a few reasons why reflecting on your year might be a good idea:

- You're essentially using System 2 to learn from the general unarticulated patterns of data gathered by System 1. This can help reduce possible Garfield errors.
- According to Dr. Jordan Peterson, written reflection also transforms the way the brain processes your experiences, as neural activity shifts from the emotional, stress-related parts of your brain (amygdala) to the more logical and detatched parts of your brain (pre-frontal cortex).
- It's always great to have an increase in self-knowledge. You are an individual after all, might as well get to know yourself better.

I struggled with clearly categorising the questions, but here's my attempt nonetheless.

**Looking at the stuff that has shapes you over the last year:**

365 days is a long time. The purpose of these questions is to become more aware of the experiences and ideas that have influenced you and the person you've become in the last year.

- What were some significant experiences that occurred in the last year? Describe 3–5. How did they change your perceptions and actions? Were they good or bad?
- What new things did you discover this year? (Books, Ideas, People) Which ones influenced you the most? How?
- What did I encounter that was unexpected—Which surprised me?
- What did I continue to enjoy this year?
- What did I continue to work on this year?
- What new priorities have emerged for you this year that you are committed to honoring? Which old ones would you like to reaffirm. Write your WHY next to each one.

**Learning from the good times and the bad times:**

- Make a list of 3–5 things that went really well in the last year. Why do you think these were successes? What stopped them from being catastrophes?
- Make a list of 3–5 things that went horribly wrong this year. Why do you think this occured? What might have led to a different outcome? What actionable changes might you make?

- What was painful? What caused this pain? List 5 examples.
- What were the largest obstacles I faced? How did I deal with them?
- What small things and big things should I be most grateful for from this year?
- *What were your worst decisions of the year? Why did you make them? What is the biggest "unfinished business" of the year and what can you do about that?*
- *What do I regret the most?*
- *What was the toughest experience?*
- *When was I at my worst?*
- What were the vices that I indulged in? Am I willing to write these off?
- *What habits am I proud to have cultivated? How can I do more of this?*
- *What beliefs got in the way last year? What are some examples?*
- What were the 20% of activities, people, and pursuits that gave me 80% of the value in my life? What 80% of things got in the way?

**Comparing Reality to Intentions:**

- What priorities did I honor well in 2017?
- What were my priorities this year—Theoretically?
- What were my actual priorities this year (as revealed by your actions—Look at your calendar). What did I turn my back on? Truly enact and commit to?
- *Which goals from the previous year did I achieve?*
- *What did I improve at?*
- *Where did I worsen or stagnate?*
- What important things were there that I should have done but didn't do? AKA What have I been avoiding?
- Why did you do what you did?
- *Did my actions show me to be a good person?*
- *How much time did I spend on things that were consequential/stuff I cared about?*
- Did I contribute to others? If not, how can I enable myself to be better in the future?…
- How did I help those around me?
- Who must I thank for being amazing this year? When will I do it?

That's it. Once again, I've made a neat tool for answering these questions in an interactive way, so it might be worth trying it.

Also, this is my first post here, so any feedback would be great. I've been thinking of sharing some summaries/reviews of books/articles clustered around certain topics IE history, psychology etc. Would that be okay?

All the best for the year ahead BTW :D

# Why did everything take so long?

One of the biggest intuitive mysteries to me is how humanity took so long to do anything.

Humans have been ‘behaviorally modern’ for about 50 thousand years. And apparently didn’t invent, for instance:

- rope until 28 thousand years ago.
- the wheel until at least 4000BC
- writing until 3000BC
- woodblock printing until 200AD

This kind of thing seems really weird introspectively, because it is hard to imagine going a whole lifetime in the wilderness without wanting something like rope, or going a whole day wanting something like rope without figuring out how to make something like rope. Yet apparently people went for about a thousand lifetimes without that happening.

Some possible explanations:

1. Inventions are usually more ingenious than they seem. LiveScience argues that it took so long to invent the wheel because “The tricky thing about the wheel is not conceiving of a cylinder rolling on its edge. It’s figuring out how to connect a stable, stationary platform to that cylinder.” I feel like that would explain why it took a month rather than a day. But a couple of thousand lifetimes?
2. Knowing what you are looking for is everything. If you sat a person down and said, “look, how do you attach a stationary platform to a rolling thing?” they could figure it out within a few hours, but if you just give them the world, they don’t think about whether a stationary platform attached to a rolling thing would be useful, so “how do you attach a stationary platform to a rolling thing” doesn’t come up as a salient question for a couple of thousand lifetimes.
3. Having concepts in general is a big deal, and being an early human who had never heard of any invention was a bit like being me when I’m half asleep.
4. Everything is always mysteriously a thousand times harder than you might think. Consider writing a blog post. Why haven’t I written a blog post in a month?
5. Others?

# TSR #6: Strength and Weakness

***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 #6: Strength and Weakness

SUMMARY

- Soviet Marshal Zhukov had his work cut out for him, and did a pretty stellar job.
- Having a clear understanding of your strengths and weakness gives insight into how to design your ops.
- Ways ops (operations) typically fail:
    - They aren't playing to your strengths
    - Don't acknowledge/mitigate weakness
    - Not understanding your s and w
    - You created a "platonic" ops that "should" work but doesn't
- "A commander must not be afraid of fighting under unfavourable circumstances."
- When looking to develop and install operations successfully, things typically don't change *quickly*, but they do change. You should periodically re-assess where you're at, so you don't fail to capitalize on gains or mitigate emerging weaknesses.
- Things have to actually work, so make sure you're dealing with the non ideal aspects of your current situation.

It's really useful to see examples of someone using principles in action. Reading about General Zhukov expertly navigating the strengths and weaknesses of the Soviets and Nazis is a great way to learn. Similarly, I really like Yudkowsky's coming of age story in the sequences. Observing the steps of someone's mind as they make mistakes or do things well is very informative.

I like the points Sebastian has made about reasons operations often fail, and I think it gives a decent guide on how to kick off a post mortem on plans and ops that you've made which haven't worked out.

Most often, my ops fail because I don't sufficiently break down the task I'm trying to operationalize. I've noticed that the more granular an ops is, the easier it is to not give into resistance in the moment. It's easier for me to "take off the covers, swing my legs down, turn off the alarm that is set to ring at the same time everyday without me having to reset it" than it is to "wake up on time". This sort of failure mode seems to be subset of "platonic ops that can't handle the real world".

Over the summer I made a new system for keeping goals and intentions in mind for my weekly planning. That system was used approximately zero times. Looking back, I see that even thought I had designed the system intending it to "operationalize" my planning, it was basically something that could only run on a constant input of willpower. No good. I'm leaning more and more towards thinking that willpower is almost never the answer to, "How should I make sure this gets done?"

I also really like the point Sebastian made about having to continually reasses the terrain. It’s easy for changes in the terrain to make your old plan obsolete, and you have to switch gears if you want to achieve what you’re after.

So, here are some useful questions to find answers to:

- What are ops you’ve tried that have failed in the past because they were too complicated or unrealistic? Have you iterated on them since?
- Are there any habits or ops you’re currently engaged in which might need to be updated based on the fact that the terrain has changed?
- Are the atomic components of your ops basic enough as to withstand the bad times? Do you need to further break down actions into smaller pieces?

# Quick thoughts on empathic metaethics

*Years ago, I wrote an unfinished sequence of posts called "No-Nonsense Metaethics." My last post, Pluralistic Moral Reductionism, said I would next explore "empathic metaethics," but I never got around to writing those posts. Recently, I wrote a high-level summary of some initial thoughts on "empathic metaethics" in section 6.1.2 of a report prepared for my employer, the Open Philanthropy Project. With my employer's permission, I've adapted that section for publication here, so that it can serve as the long-overdue concluding post in my sequence on metaethics.*

In my previous post, I distinguished "austere metaethics" and "empathic metaethics," where austere metaethics confronts moral questions roughly like this:

> Tell me what you mean by 'right', and I will tell you what is the right thing to do. If by 'right' you mean X, then Y is the right thing to do. If by 'right' you mean P, then Z is the right thing to do. But if you can't tell me what you mean by 'right', then you have failed to ask a coherent question, and no one can answer an incoherent question.

Meanwhile, empathic metaethics says instead:

> You may not know what you mean by 'right.' But let's not stop there. Here, let me come alongside you and help decode the cognitive algorithms that generated your question in the first place, and then we'll be able to answer your question. Then we can tell you what the right thing to do is.

Below, I provide a high-level summary of some of my initial thoughts on what one approach to "empathic metaethics" could look like.

Given my metaethical approach, when I make a “moral judgment” about something (e.g. about which kinds of beings are moral patients), I don’t conceive of myself as perceiving an objective moral truth, or coming to know an objective moral truth via a series of arguments. Nor do I conceive of myself as merely expressing my moral feelings as they stand today. Rather, I conceive of myself as making a conditional forecast about what my values would be if I underwent a certain “idealization” or “extrapolation” procedure (coming to know more true facts, having more time to consider moral arguments, etc.).[1]

Thus, in a (hypothetical) "extreme effort" attempt to engage in empathic metaethics (for thinking about *my own* moral judgments), I would do something like the following:

1. I would try to make the scenario I'm aiming to forecast as concrete as possible, so that my brain is able to treat it as a genuine forecasting challenge, akin to participating in a prediction market or forecasting tournament, rather than as a fantasy about which my brain feels "allowed" to make up whatever story feels nice, or signals my values to others, or achieves something else that isn't *forecasting accuracy*.[2] In my case, I concretize the extrapolation procedure as one involving a large population of copies of me who learn many true facts, consider many moral arguments, and undergo various other experiences, and then collectively advise me about what I should value and why.[3]

2. However, I would also try to make forecasts I can actually check for accuracy, e.g. about what my moral judgment about various cases will be 2 months in the future.
3. When making these forecasts, I would try to draw on the best research I've seen concerning how to make accurate estimates and forecasts. For example I would try to "think like a fox, not like a hedgehog," and I've already done several hours of probability calibration training, and some amount of forecasting training.[4]
4. Clearly, my current moral intuitions would serve as one important source of evidence about what my extrapolated values might be. However, recent findings in moral psychology and related fields lead me to assign more evidential weight to some moral intuitions than to others. More generally, I interpret my current moral intuitions as data generated partly by my moral principles, and partly by various "error processes" (e.g. a hard-wired disgust reaction to spiders, which I don't endorse upon reflection). Doing so allows me to make use of some standard lessons from statistical curve-fitting when thinking about how much evidential weight to assign to particular moral intuitions.[5]
5. As part of forecasting what my extrapolated values might be, I would try to consider different processes and contexts that could generate alternate moral intuitions in moral reasoners both similar and dissimilar to my current self, and I would try to consider how I feel about the the "legitimacy" of those mechanisms as producers of moral intuitions. For example I might ask myself questions such as "How might I feel about that practice if I was born into a world in which it was already commonplace?" and "How might I feel about that case if my built-in (and largely unconscious) processes for associative learning and imitative learning had been exposed to different life histories than my own?" and "How might I feel about that case if I had been born in a different century, or a different country, or with a greater propensity for clinical depression?" and "How might a moral reasoner on another planet feel about that case if it belonged to a more strongly r-selected species (compared to humans) but had roughly human-like general reasoning ability?"[6]
6. Observable patterns in how people's values change (seemingly) in response to components of my proposed extrapolation procedure (learning more facts, considering moral arguments, etc.) would serve as another source of evidence about what my extrapolated values might be. For example, the correlation between aggregate human knowledge and our "expanding circle of moral concern" (Singer 2011) might (very weakly) suggest that, if I continued to learn more true facts, my circle of moral concern would continue to expand. Unfortunately, such correlations are badly confounded, and might not provide much evidence.[7]
7. Personal facts about how my own values have evolved as I've learned more, considered moral arguments, and so on, would serve as yet another source of evidence about what my extrapolated values might be. Of course, these relations are likely confounded as well, and need to be interpreted with care.[8]

---

**1.** This general approach sometimes goes by names such as "ideal advisor theory" or, arguably, "reflective equilibrium." Diverse sources explicating various extrapolation procedures (or fragments of extrapolation procedures) include: Rosati (1995); Daniels (2016); Campbell (2013); chapter 9 of Miller (2013); Muehlhauser & Williamson (2013); Trout (2014); Yudkowsky's "Extrapolated volition (normative moral theory)" (2016); Baker (2016); Stanovich (2004), pp. 224-275; Stanovich (2013).

**2.** For more on forecasting accuracy, see this blog post. My use of research on the psychological predictors of forecasting accuracy for the purposes of doing moral

philosophy is one example of my support for the use of "ameliorative psychology" in philosophical practice — see e.g. Bishop & Trout (2004, 2008).

**3.** Specifically, the scenario I try to imagine (and make conditional forecasts about) looks something like this:

1. In the distant future, I am non-destructively "uploaded." In other words, my brain and some supporting cells are scanned (non-destructively) at a fine enough spatial and chemical resolution that, when this scan is combined with accurate models of how different cell types carry out their information-processing functions, one can create an executable computer model of my brain that matches my biological brain's input-output behavior almost exactly. This whole brain emulation ("em") is then connected to a virtual world: computed inputs are fed to the em's (now virtual) signal transduction neurons for sight, sound, etc., and computed outputs from the em's virtual arm movements, speech, etc. are received by the virtual world, which computes appropriate changes to the virtual world in response. (I don't think anything remotely like this will ever happen, but as far as I know it is a *physically possible* world that can be described in some detail; for one attempt, see Hanson 2016.) Given functionalism, this "em" has the same memories, personality, and conscious experience that I have, though it experiences quite a shock when it awakens to a virtual world that might look and feel somewhat different from the "real" world.
2. This initial em is copied thousands of times. Some of the copies interact inside the same virtual world, other copies are placed inside isolated virtual worlds.
3. Then, these ems spend a very long time (a) collecting and generating arguments and evidence about morality and related topics, (b) undergoing various experiences, in varying orders, and reflecting on those experiences, (c) dialoguing with ems sourced from other biological humans who have different values than I do, and perhaps with sophisticated chat-bots meant to simulate the plausible reasoning of other types of people (from the past, or from other worlds) who were not available to be uploaded, and so on. They are able to do these things for a very long time because they and their virtual worlds are run at speeds thousands of times faster than my biological brain runs, allowing subjective eons to pass in mere months of "objective" time.
4. Finally, at some time, the ems dialogue with each other about which values seem "best," they engage in moral trade (Ord 2015), and they try to explain to me what values they think I should have and why. In the end, I am not forced to accept any of the values they then hold (collectively or individually), but I am able to come to much better-informed moral judgments than I could have without their input.

For more context on this sort of values extrapolation procedure, see Muehlhauser & Williamson (2013).

**4.** For more on forecasting "best practices," see this blog post.

**5.** Following Hanson (2002) and ch. 2 of Beckstead (2013), I consider my moral intuitions in the context of Bayesian curve-fitting. To explain, I'll quote Beckstead (2013) at some length:

> Curve fitting is a problem frequently discussed in the philosophy of science. In the standard presentation, a scientist is given some data points, usually with an independent variable and a dependent variable, and is asked to predict the values of the dependent variable given other values of the independent variable.

> Typically, the data points are *observations*, such as "measured height" on a scale or "reported income" on a survey, rather than true values, such as height or income. Thus, in making predictions about additional data points, the scientist has to account for the possibility of error in the observations. By an error process I mean anything that makes the observed values of the data points differ from their true values. Error processes could arise from a faulty scale, failures of memory on the part of survey participants, bias on the part of the experimenter, or any number of other sources. While some treatments of this problem focus on predicting observations (such as measured height), I'm going to focus on predicting the true values (such as true height).
>
> …For any consistent data set, it is possible to construct a curve that fits the data exactly… If the scientist chooses one of these polynomial curves for predictive purposes, the result will usually be *overfitting*, and the scientist will make worse predictions than he would have if he had chosen a curve that did not fit the data as well, but had other virtues, such as a straight line. On the other hand, always going with the simplest curve and giving no weight to the data leads to *underfitting*…
>
> I intend to carry over our thinking about curve fitting in science to reflective equilibrium in moral philosophy, so I should note immediately that curve fitting is not limited to the case of two variables. When we must understand relationships between multiple variables, we can turn to multiple-dimensional spaces and fit planes (or hyperplanes) to our data points. Different axes might correspond to different considerations which seem relevant (such as total well-being, equality, number of people, fairness, etc.), and another axis could correspond to the value of the alternative, which we can assume is a function of the relevant considerations. Direct Bayesian updating on such data points would be impractical, but the philosophical issues will not be affected by these difficulties.
>
> …On a Bayesian approach to this problem, the scientist would consider a number of different hypotheses about the relationship between the two variables, including both hypotheses about the phenomena (the relationship between X and Y) and hypotheses about the error process (the relationship between observed values of Y and true values of Y) that produces the observations…
>
> …Lessons from the Bayesian approach to curve fitting apply to moral philosophy. Our moral intuitions are the data, and there are error processes that make our moral intuitions deviate from the truth. The complete moral theories under consideration are the hypotheses about the phenomena. (Here, I use "theory" broadly to include any complete set of possibilities about the moral truth. My use of the word "theory" does not assume that the truth about morality is simple, systematic, and neat rather than complex, circumstantial, and messy.) If we expect the error processes to be widespread and significant, we must rely on our priors more. If we expect the error processes to be, in addition, biased and correlated, then we will have to rely significantly on our priors even when we have a lot of intuitive data.

Beckstead then summarizes the framework with a table (p. 32), edited to fit into LessWrong's formatting:

- Hypotheses about phenomena
    - *(Science)* Different trajectories of a ball that has been dropped

    - *(Moral Philosophy)* Moral theories (specific versions of utilitarianism, Kantianism, contractualism, pluralistic deontology, etc.)
- Hypotheses about error processes
    - *(Science)* Our position measurements are accurate on average, and are within 1 inch 95% of the time (with normally distributed error)
    - *(Moral Philosophy)* Different hypotheses about the causes of error in historical cases; cognitive and moral biases; different hypotheses about the biases that cause inconsistent judgments in important philosophical cases
- Observations
    - *(Science)* Recorded position of a ball at different times recorded with a certain clock
    - *(Moral Philosophy)* Intuitions about particular cases or general principles, and any other relevant observations
- Background theory
    - *(Science)* The ball never bounces higher than the height it started at. The ball always moves along a continuous trajectory.
    - *(Moral Philosophy)* Meta-ethical or normative background theory (or theories)

**6.** For more on this, see my conversation with Carl Shulman, O'Neill (2015), the literature on the evolution of moral values (e.g. de Waal et al. 2014; Sinnott-Armstrong & Miller 2007; Joyce 2005), the literature on moral psychology more generally (e.g. Graham et al. 2013; Doris 2010; Liao 2016; Christen et al. 2014; Sunstein 2005), the literature on how moral values vary between cultures and eras (e.g. see Flanagan 2016; Inglehart & Welzel 2010; Pinker 2011; Morris 2015; Friedman 2005; Prinz 2007, pp. 187-195), and the literature on moral thought experiments (e.g. Tittle 2004, ch. 7). See also Wilson (2016)'s comments on internal and external validity in ethical thought experiments, and Bakker (2017) on "alien philosophy."

I do not read much fiction, but I suspect that some types of fiction — e.g. historical fiction, fantasy, and science fiction — can help readers to temporarily transport themselves into fully-realized alternate realities, in which readers can test how their moral intuitions differ when they are temporarily "lost" in an alternate world.

**7.** There are many sources which discuss how people's values seem to change along with (and perhaps in response to) components of my proposed extrapolation procedure, such as learning more facts, reasoning through more moral arguments, and dialoguing with others who have different values. See e.g. Inglehart & Welzel (2010), Pinker (2011), Shermer (2015), and Buchanan & Powell (2016). See also the literatures on "enlightened preferences" (Althaus 2003, chs. 4-6) and on "deliberative polling."

**8.** For example, as I've learned more, considered more moral arguments, and dialogued more with people who don't share my values, my moral values have become more "secular-rational" and "self-expressive" (Inglehart & Welzel 2010), more geographically global, more extensive (e.g. throughout more of the animal kingdom), less person-affecting, and subject to greater moral uncertainty (Bykvist 2017).

# Maps vs Buttons; Nerds vs Normies

It took me the longest time to see through the illusion that was "rational" discussion.

The setting: A friend and I exchange inquiries about each other's beliefs about X.

The result: the friend would give a number of answers that allowed me to piece together their view of X. An astounding percentage of the time, even a majority, perhaps, one of their answers would contradict another. So, I point this out. My friend, after hearing an explanation, agrees it is a contradiction. We move on, me convinced that I had contributed to their work (the work I assumed that others, not just myself, engaged in: namely, piecing together as accurate a model of reality as possible), only to find, weeks or months later, that on discussing X again, they still gave the same contradictory answers as before, with apparently no memory of our past discussion.

Or discussions. This pattern has repeated up to 4 times (that I've bothered to keep track), my interlocutor agreeing with my corrections, and then showing no sign of having even heard of them down the road.

My whole above description of "the result" is faulty in a subtle way, one that undermines the entirety of it. Here's what was really happening, expressed from my friend's perspective.

The (real) result: A friend and I were discussing X. They showed respect for me by asking me to express myself on it several times. At some point, they dared to point out, rather like a pedantic schoolboy, that there was a sort of inconsistency between my expressions and the forms we studied in textbooks on logic. They were not rude, though, so I graciously acknowledged the rather dry, uninteresting observation, after which we happily continued our conversation. I had thought for a moment that they wished to challenge me, but all seemed to be rather well-resolved, so I forgot the particulars of the incident, only taking with me the general, updated state of our relationship, including, for example, the points that my expressions were generally respected, but that my friend was willing to make some small challenges if they felt like it, though also willing to move on after I was gracious in response.

Or something like this, perhaps. And this key difference I call the map vs the button idea.

I, and other nerds, are conceptual cartographers, attempting to piece together a map of reality. I've often wondered how I seem to update my map so consistently, since I make no specific effort to remember any alterations or additions I deem worth making. Shouldn't I be forgetting them? Maybe I should dedicate some time to writing down and reviewing these precious insights!

But I've since realized that my map is very personal to me. I am well-acquainted with its pieces, which are interconnected in many ways. So long as I swap out one piece at a time, while maintaining mostly the same interconnections with the replacement piece, I seem to remember such changes very easily.

When you ask me what my beliefs on some subject are, I consult my mental map, and I produce for you, on the fly(!), a description of what the map says. How do I answer your inquiry? I deduce my answer to your question by referring to my map.

Two points:

1. This produces conceptually consistent answers, as consistent as my map is, which is to say, very, since I constantly compare its pieces against each other.
2. My expressions, my answers to questions, are often clunky and unwieldy, since I've constructed them on the fly. If you're like me, you might wonder why this should make my answers any clunkier than anyone else's, for surely this is what everyone does, right? Herein lies a clue to the great insight.

Now the contrast. Others are not making maps. I think of them rather as like a big wall of buttons. Push a button, and a slip of paper comes out. When I ask such a button-type person a question, I am pushing a certain button, and they, without even thinking, spit out the coresponding piece of paper (there is some relatedness here to the idea of "cached thoughts" as Eliezer Yudkowsky spoke of them. You could well say that these people are mostly just vast collections of cached thoughts).

Where do they spend their effort, if not on map-making? On the following:

1. Reading your reaction to their slips of paper (their answers).
2. Re-writing and refining the writing on the papers. The higher your status, the more your reaction to their slips of paper compels them to change them.

Some important differences that flow naturally out from these two different methods and focuses:

1. Mapmaking nerds give more consistent answers than button people, since they view their map as a whole, each part being compared against the others, whereas the slips of paper are not judged (even in part) by how well they match other slips of paper, but rather they are judged according to the reactions that they evoke in the audience.
2. Button people give much more eloquent, refined, and most of all, *socially advantageous* answers than mapmakers, and do so with less apparent difficulty or searching for words, analogies, descriptions. This improvement in delivery comes from the fact that they don't have any thinking to do on the fly; they're just repeating the same thing as always (have you ever heard people who always give, word-for-word, the same commentary or response to some particular issue which may be brought up?), so their answer comes out smoothly and immediately. (Indeed, if a question does not *exactly* ask something they have a prepared response for, they may well give one of their rehearsed lines, anyway, considering the cost of imprecisely answering less than the cost of having to stop and, you know, *think* about how to answer. Less chance to err, you see) Also, since the answer has, effectively speaking, been "practiced" many times in the past, and has been refined over time as ideas come to them, or as they seek to ameliorate negative reactions, there is greater eloquence and ease of presentation. (It occurs to me that this may also explain why the "best" answers according to this kind of system, especially when addresssing highly controversial issues, are sometimes as unclear and meaningless as possible. Suppose a piece of your answer offends people A, so you take it out. Then, another piece offends people B, so you transform it into something inoffensive, and so on. Eventually, you end up with an answer that gives everyone generally positive feelings, without pissing anybody off, but which is so devoid of content that not only was it *not* deduced from a map of reality, but the whole method of deducing replies by consulting models could never produce it, so incoherent and meaningless is the expression! That is, a nerd, looking at their map and

> reporting what they saw, would never produce such a statement, because it is not the kind of statement that comes from trying to describe how something works or how it is. Rather, these meaningless niceties are the kind of statement that come from trying to appease and impress people with the content and presentation of your social signals. Statements which embody these social goals to this extreme degree sound *so* completely different from the ones that nerds use to communicate actual ideas, that the illusion is broken, and instead of thinking their interlocutor is reporting on their personal map when they're not (as was my common error (see 1st paragraph)), a nerd is likely to just give a blank stare and ask for clarification (I feel this is whole aside is somewhat unclear, but if you're me, or like me, I hope it has a ring of familiarity to it), receiving only more incoherent niceties in response (since there *is* no content to be brought into focus) until they or the interlocutor tires and abandons the project).

Belief(1) shall refer to the kind of thing a nerd says when asked what they believe. It is as accurte a report of their model of reality as they can give on the fly.

Belief(2) shall refer to the kind of answer a normal, social person gives when asked. It is as nice (to their in-group, anyway) and as impressive a statement on the subject as they so far know how to give. It's very much like whipping out the verbal version of a pretty bauble to gain *ooh*'*s* and *aah*'*s (*attention and status).

And so I say, belief(1) =/= belief(2).

In a very real sense, normal people just don't even *have* beliefs(1)! More precisely, if we're talking about things where they don't need to be factually right in order to prosper, if we're talking about politics, religion, philosophy, etc., they don't have beliefs(1) at all.

At the same time, they sustain a convincing illusion of having beliefs(1) (not that that's on purpose or anything; thinking in terms of beliefs(2) is their natural instinct), because when you say "what's your belief on gender equality", they immediately produce a nice shiny answer that definitely *sounds* like it's saying something about the nature of reality.

The natural but mistaken thing to do is treat what is an automatic, unthinking, knee-jerk, instinctive attempt to impress and to signal which groups they owe fealty to as instead an attempt to describe some part of reality, merely because all of the words in their response sound like a description of a part of reality. That's the error.

So, nerds talk in beliefs(1), normal people talk in beliefs(2), and the result is that both sides commit mutual faux pas and talk past each other. Nerds are constantly embarrassing themselves according to social rules, and normal people don't have any ideas worth listening according to nerd rules.

The specifics of these errors can be deduced once you realize that each side sees their own methodology as so natural that they assume the *other* side is also using it, but are doing so incompetently. So nerds think normal people are failed nerds, rather than successful normal people, while normal people think nerds are failed normal people, rather than successful nerds.

Normal people think nerds are trying to signal and impress (and failing pathetically and amusingly) and nerds think normal people are trying to make models of the world,

and epic fail so hard as to be incapable of discussing any subject beyond four sentences without contradicting themselves. And so on.

If you're a nerd, you might read all this and think I'm being hard on normal people (how can you say such awful things about them as that they're not logically consistent and that they don't ponder before answering questions?), while if you're a normal person reading this (haha, jk), you might think I'm awful hard on nerds (how can you say such mean things as that they don't care what others think and are incapable of properly expressing themselves?). This phenomenon occurs again because both sides are judging everybody by their own standard, not recognizing that others have other standards and succeed very well by them.

One last note. I've spoken as if people are in one camp or the other, but it's really more of a spectrum. Even more precise, it might be correct to say that people are on a spectrum of nerdy/normal for *each specific topic* they care about. Frankly, very few humans manage to be nerdy about politics and religion, even "nerds." People who are nerdy in many ways suddenly turn into political animals, social thinkers, when you bring up anything controversial. And a normal person might well become nerdy and actually try to learn how some thing *really* works if they happen to be interested in it for some reason.

# Methods of Phenomenology

This is a linkpost for https://mapandterritory.org/methods-of-phenomenology-e2f936651ff

*NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.*

In the previous post we looked at phenomenology and the thinking that motivates it. We saw that it is based on taking a naive, skeptical, beginner's view to asking "why?" and choosing to address the question using only knowledge we can obtain from experience. We also saw that experience is intentional: that it is directed from subject to object and forms an inseparable, arity 3 relation called a phenomenon. Taken together this gave us a feel for the shape of phenomenological philosophy and allowed us to glimpse some of the consequences of taking this view seriously.

We now have the context to begin exploring phenomenology's details, and the first detail to explore is the methods of exploration themselves because phenomenology is highly integrative and our assumptions—that we know the world only through experience and that experience is intentional—determine what sorts of methods we can use. Thus if we are to approach phenomenology we must first gain some familiarity with its practices.

Now if I'm honest there is only really one method of phenomenology—the phenomenological reduction—but that's a bit like telling you that the only method of decision making is Bayes's Theorem: in a sense it's true, but it's not likely to help you understand anything in the same way that telling you that you are already enlightened doesn't make you enlightened. To explain how phenomenologists think, it's more useful to talk about a broad base of specific methods and through them approach the reduction proper. Luckily, there are many methods, and you are almost certainly already familiar with several of them, so let's take a look!

## Science

It might seem a little surprising to you that science is a method of phenomenology since, historically, phenomenology emerged in part from Husserl noticing the inadequacy of natural science for addressing conscious experience, and, in the latter-half of the 20th century, phenomenologically inspired thinkers in the post-modern movement sometimes took anti-scientific stances. But Husserl viewed science as part of phenomenology, and some consider phenomenology a science of consciousness, so there's more to the relationship between science and phenomenology than some surface level antagonism.

Like phenomenology, science starts from a place of empiricism—the idea that knowledge is obtained through experience and observation. To the extent that science is systematized empiricism, phenomenology is a science, but "science" usually refers to a more specific practice of the scientific method with certain standards of evidence that phenomenology doesn't always hold itself to. In particular, science considers only phenomena where subjects and intentionality can be ignored because the objects and the experiences of them remain relatively unchanged between subjects. We call such truncated, replicable experiences of objects objective or natural phenomena.

One of the foundational issues of science is to decide exactly what the criteria for objective phenomenon are, but generally objective phenomena are those that describe sufficiently similar experiences of objects no matter who or what the subject is, like the way a beam of light will be experienced as having the same wavelength no matter who sees it or what measures it. By limiting itself to these objective phenomena, science is able to make predictions about the world that it expects to hold for all subjects, which is to say that what is true of a few phenomena will be true of all similar phenomena. This let's us uncover patterns science calls theories that make strong predictions about the world.

As so far described, science is compatible with phenomenology and allows us to make much more confident statements about the world when objective phenomena are available than when they are not. But objective phenomena cannot always be reliably constructed. Because objective phenomena are not actually phenomena but patterns of statistical regularity observed over many phenomena, there is necessarily information lost in the creation of objective phenomena, limiting what can be known through them. This might seem like an obscure, technical issue for philosophers of science, but it has implications for when science, in the sense of understanding the world through the use of objective phenomena, is an appropriate method.

Phenomenology views science as extremely useful but sees it running aground the closer it gets to exploring topics where the intentional nature of experience matters and objective phenomena are less available, such as in the study of consciousness. Thus science is a great method for exploring questions of physics, chemistry, and biology; pretty good for studying economics and archeology; workable in psychology and anthropology; and of limited direct usefulness in philosophy and philology. For those topics where science cannot cover all the epistemology ground, other phenomenological methods are necessary.

**An Aside on Scientism, Irrationality, and Their Kin**

Now I'd rather not have to write this part but I suspect some readers may be upset at me for presenting the relationship between phenomenology and science as prosaic. The technical issue of determining how much we can figure out with science alone has and does get mixed up with all sorts of discussions about other things, so I think it's worth saying a few words about this to at least acknowledge the issue and direct you to additional reading if this topic is of interest.

Humans are political animals, so when there is disagreement on something it often sparks or gets sucked into a larger battle between groups. One of these battles is along a dimension we might call "rationality" between those who value the modern worldview and those who don't. The details get complicated, but you can basically imagine it as if there were two political parties vying for control of a country, the Pro-Rationality Party and the Anti-Rationality Party, and it's into this milieu that phenomenology and science are thrown.

The Pros claim science for their own, so the Antis reject it. Phenomenology says science is useful for understanding many things but not literally all things, so the extremists on the Pro side reject phenomenology for being "impure". The Anti side then takes phenomenology in and plays up the limits-of-science thing while downplaying the usefulness-of-science part. As a result phenomenologists more often find themselves having to defend their ideas against material realism, scientism, and other ideas on the Pro side and less against irrationality, mysticism, and other incompatible positions on the Anti side. This creates a skewed picture that implies

phenomenology is anti-science by association, and it doesn't help that some phenomenologists, being humans, may actually take up sides in this debate.

But ultimately the Pro/Anti battle is more about how humans relate to ideas than the ideas themselves, and methods like debate magnify this confusion. Thankfully, phenomenologists and other philosophers have an alternative to debate that functions better at collaborative truth seeking: the dialectic.

## Dialectic

Philosophers have a special way of talking to each other in good faith that cuts to the heart of their disagreements. Once they find these disagreements they can build toward mutual understanding and possible agreement. We generically call this process dialect and I've written about it before:

**Debate and Dialectic**
*Election season is over in the US, but folks are still talking about how divided political conversation is. We hear…*mapandterritory.org

In case you don't want to click the link, the short version is that the dialectical method is to consider a position or idea, called the thesis, find something that contradicts it, an antithesis, and then try to find a synthesis of thesis and antithesis that sublimates/overcomes them with new understanding. That new understanding becomes a new thesis, and the process repeats until it converges to consensus or diverges to logical inconsistency.

Dialectic differs from debate so long as the antithesis and synthesis are formed in good faith. With good faith, dialectic can get at the heart of a dispute to find agreement, but with bad faith dialectic devolves into debate and drives a wedge between people and ideas. Since phenomenology gives primacy to phenomena, including those non-objective experiences we might call "subjective", it strongly encourages good faith and is able to make extensive use of dialectic as a tool for building understanding.

There is not always disagreement or apparent contradiction to power a dialectic, though. In those cases phenomenologists must explore the world in other ways, and more closely examining the phenomena themselves often yields dividends.

## Hermeneutics

When we write, speak, or otherwise communicate, we engage in an act of creating phenomena for others by giving them objects to experience. We can try to anticipate how they will experience these objects—to predict the phenomena our audiences will find themselves subject to when they read our writing, hear our words, or see our art—but there will inevitably be variation in their experiences. This means that for all experiences of the same object there will be different experiences had by different subjects. This opens up the opportunity to compare and study the differences in experiences, and we call this study hermeneutics.

Although technically it is possible to perform hermeneutics on phenomena of non-conscious subjects, we generally consider that practice a part of applied science or engineering, so hermeneutics generally refers to the process of interpreting the experiences of conscious subjects. Originally hermeneutics primarily focused on

interpretation of sacred experiences, especially of messages believed to have been sent by the gods, but Heidegger generalized the notion within phenomenology to interpretation of experience and developed the hermeneutic circle as his primary philosophical technique. Philologists then mixed Heidegger's philosophical hermeneutics with their own methods and developed techniques we now think of as literary criticism, historical analysis, and other methods of critical study in the humanities.

I think of hermeneutics as a kind of meditation on the experiences of others where people report their experiences and we think on those reports to create our own experiences of them. We only have access to our own experiences, but from our experiences we can reason about the world that made possible the experiences of others and so gain partial, indirect knowledge of objects of experience we never experienced ourselves. In this sense hermeneutics is what we do whenever we read a book, listen to a friend talk, or empathize with the experiences of others.

We can similarly think of meditation as hermeneutical analysis of our own experiences, but this would be selling meditation short because, unlike when analyzing the reported experiences of others, we are the subjects of our own experiences and can, at least in theory, know more about them. Turning this theory to practice is not easy, though, so meditation is a method of phenomenological epistemology worth exploring on its own.

## Meditation

"Meditation" is a word with a lot of meanings. In one sense it means focused thinking on a topic, and you might say my writings are often meditations of this sort. There's another sense in which meditation is the practice of entering trances and other altered states of consciousness, possibly associated with spiritual experiences, and while this is interesting because it may produce qualia not otherwise generated, it's not a phenomenological technique so much as a source of capta. Instead the sense in which we care about meditation is as a method of cultivating awareness of the world and our interactions within it so that we may learn everything we can from our experiences.

There are many specific meditative practices that can serve phenomenological purposes. For example, the meditation of early phenomenologists was heavily influenced by yogacara, I practice zazen, and any technique that teaches the ability to observe phenomena without interpretation will work. The key is learning to withhold judgement so that, as much as possible, the world may be seen as it is. From gaining such a clear picture of the world we may start our naive, skeptical, beginner's investigation of it.

Being skeptical, it's fair to ask how much value we can derive from meditation. After all, psychology is littered with disproved theories that drew much of their evidence from introspection, so there seems reason to be suspect of anyone claiming knowledge solely based on their own experiences. But just as science abandoned those theories when their evidence did not reproduce, the phenomenological framework similarly does not ask you to accept the evidence provided by others (or yourself!) blindly. If someone reports an experience that seems false to you in some way, you should try to understand it, and if you desire to know more you should try meditating on similar experiences yourself to see what you learn. If you get different results than others, that can be a starting point for dialectic and hermeneutics.

Thus it's important to be clear that meditation is like science, dialectic, and hermeneutics in that it does not stand on its own. Meditation cannot give us perfect knowledge even as it helps us to approach the limits of our knowledge imposed by the intentional nature of experience. But how close can we get to those limits? Husserl believed it was possible to get so close as to feel yourself transcending them, but any such feeling of transcendence must itself be an experience that can be suspended and examined, so it seems at best we can reach an equilibrium of continuously experiencing the experience of experiencing experience. To make sense of such deeply self-referential phenomena, Husserl developed the phenomenological reduction, the foundational method of phenomenology.

## The Phenomenological Reduction

All phenomenological methods are expressions of the phenomenological reduction. They're not like this because they were designed this way: most phenomenological methods predate the idea of phenomenology itself. Instead, the phenomenological reduction is the core movement available to us as we explore the world via phenomena, and so all other methods are naturally expressions of it. That we do not always use the naked reduction directly reflects the difficulty of carrying out the reduction in full.

The reduction is not very easy to describe, either. It consists of a single movement with two motions—epoche and epistrophe. "Epoche" is the Greek word Husserl used to refer to the process of suspending, stepping back from, or bracketing an experience so that it may be examined, and epistrophe is the dual or reverse process of epoche where we return, reintegrate, or reduce our understanding back from suspension. Confusingly Husserl didn't use the term "epistrophe" to match "epoche" but instead referred to epistrophe as "the reduction proper" (German: *das eigentliche Reduzieren*, "the reduction in its own light") based on the original Latin meaning of *reducere* from *re-* meaning "back" or "again" and *ducere* meaning "lead" or "bring". Given the confusion this invites both because it gives too similar names to the method and one of its motions and because "reduction" is now philosophically cognate with reductionism, I choose to use "epistrophe" instead.

Notice that I called epoche and epistrophe motions and not steps or parts. This is intentional because the reduction is a complete movement where one motion naturally follows the other. You might think of epoche as breathing in, epistrophe as breathing out, and reduction as breathing: you have to breathe in and breathe out to breathe, if you breathe in you necessarily breathe out, and if you breathe out you will almost certainly breath in again. Thus although we may talk about the two motions separately, they fundamentally imply one another.

To see the reduction at work, let's perform its motions on a classic example from phenomenology, seeing a cup.

We first perform epoche by suspending the action of seeing the cup to experience the phenomenon of seeing a cup as phenomenon. That is, we quote the phenomenon “I see a cup” so that we can consider it apart from our participation in it. From there we might bracket the phenomenon further to find that, for example, what we think of as a cup is actually our mind interpreting particular sensations as a cup, and those sensations are themselves further phenomena of cells in our body producing chemical-electrical signals in response to light. We can continue epoche until the phenomena we wish to examine have been bracketed or we lack the insight to see a further suspension.

We then move into epistrophe to reconstruct what we deconstructed via epoche having gained a broader perspective. Before a cup was just a cup; now we can see a cup as a construction of multiple layers of phenomena adding bits of meaning—what we might also call ontology, categorization, modeling, pattern matching, or the map—leading up to an experience of seeing a cup as a cup. Along the way we get a picture of how a cup comes to be and how it is differentiated from other things, and if we still do not see reality as clearly as we need to after this, we move back into epoche to begin the reduction again.

The cycle of epoche and epistrophe forces us to be parsimonious. If we leave in epicycles or other sorts of complex assumptions during epoche or construct them during epistrophe, we will merely find ourselves needing to further bracket our perceptions until we can explain them to ourselves. And if we repeat the reduction long enough and take our thinking far enough, we end up doing what Husserl said made the reduction “radical”: we gain gnosis of consciousness and being. We’ll return to issues of consciousness soon enough, so for now let’s wrap up by seeing how reduction connects all of our methods.

Being trained in mathematical thinking, I often engage in a formal version of the phenomenological reduction to make my thinking clear (cf. how I showed my epoche in the introduction to phenomenological AI alignment). Husserl preferred to practice radical self-meditation, as he put it, to perform the phenomenological reduction. Heidegger’s hermeneutic circle is a thin layer on top of phenomenological reduction, focused on seeing things alternately as made of parts and as wholes rather than as a movement between epoche and epistrophe. Dialectics often play out as epoche until they converge and enable epistrophe. And science most of all cherishes epoche as its method of suspending judgement to see the world as it is so that theories may be developed under formal rules of epistrophe. All, when done from a phenomenological perspective, embody the motions of the reduction and *the* method of phenomenology.

Next time we’ll begin exploring aspects of the world from a phenomenological perspective to prepare us for talking about noematological alignment in AIs. See you then!

# Happiness Is a Chore

This is a linkpost for http://squirrelinhell.blogspot.com.es/2017/12/happiness-is-chore.html

# Comments on Power Law Distribution of Individual Impact

I had a discussion online yesterday, stemming from whether you should expect to be able to identify individuals who will most shape the long term future of humanity. It was on a discussion of whether CEA should have staff work on doing this full time, and I was expecting boring comments that just expressed a political opinion about what CEA should do. However, Jan Kulveit offered some concrete models for me to disagree with, and I had a fun exchange and appreciated the chance to make explicit some of my models in this area.

With permission of all involved, I have reproduced the exchange below.

---

*Jan*:

I would be also worried. Homophily is of the best predictors of links in social networks, and factors like being member of the same social group, having similar education, opinions, etc. are known to bias selection processes again toward selecting similar people. This risks having the core of the movement be more self encapsulated that it is, which is a shift in bad direction.

Also I would be worried with 80k hours shifting also more toward individual coaching, there is now a bit overemphasis on "individual" approach and too little on "creating systems".

Also it seems lot of this would benefit from knowledge from the fields of "science of success", general scientometry, network science, etc. E.g. when I read concepts like "next Peter Singer" or a lot of thinking along the line "most of the value is created by just a few peple", I'm worried. While such thinking is intuitively appealing, it can be quite superficial. E.g., a toy model: Imagine a landscape with gold scattered in power-law sized deposits. And prospectors, walking randomly, and randomly discovering deposits of gold. What you observe is the value of gold collected by prospectors is also power-law distributed. But obviously the attempts to emulate "the best" or find the "next best" would be futile. It seems open question (worth studying) how much some specific knowledge landscape resembles this model, or how big part of the success is attributable to luck.

---

*Ben (me):*

That's a nice toy model, thanks for being so clear :-)

But it's definitely wrong. If you look at Bostrom on AI or Einstein on Relativity or Feynman on Quantum Mechanics, you don't see people who are roughly as competent as their peers, just being lucky in which part of the research space was divvied up and given to them. You tend to see people with rare and useful thinking processes having multiple important insights about their field in succession - getting many thing right that their peers didn't, not just one as your model would predict (if being right was random luck). Bostrom has looked into half a dozen sci-fi looking areas that others looked to figure out which were important, before concluding with xrisk and AI, and he

looked into areas and asked questions that were on nobody's radar. Feynman made breakthroughs in many different subfields, and his success looked like being very good at fundamentals like being concrete and noticing his confusion. I know less about Einstein, but as I understand it to get to Relativity required a long chain of reasoning that was unclear to his contemporaries. "How would I design the universe if I were god" was probably not a standard tool that was handed out to many physicists to try.

You may respond "sure, these people came up with lots of good ideas that their contemporaries wouldn't have, but this was probably due to them using the right heuristics, which you can think of as having been handed out randomly in grad school to all the different researchers, so it still is random just on the level of cognitive processes".

To this I'd say that, you're right, looking at people's general cognitive processes is really important, but I think I can do much better than random chance in predicting what cognitive processes will produce valuable insights. I'll point to Superforecasters and Rationality: AI to Zombies as books with many insights into which cognitive processes are more likely to find novel and important truths than others.

In sum: I think the people who've had the most positive impact in history are power law distributed because of their rare and valuable cognitive processes, not just random luck, and that these can be learned from and that can guide my search for people who (in future) will have massive impact.

---

*Jan:*

Obviously the toy model is wrong in describing reality: it's one end of the possible spectrum, where you have complete randomness. On the other you have another toy model: results in a field neatly ordered by cognitive difficulty, and the best person at a time picks all the available fruit. My actual claims roughly are

- reality is somewhere in between
- it is field-dependent
- even in fields more toward the random end, there actually would be differences like different speeds of travel among prospectors

It is quite unclear to me where on this scale the relevant fields are.

I believe your conclusion, that the power law distribution is all due to the properties of the peoples cognitive processes, and no to the randomness of the field, is not supported by the scientometric data for many research fields.

Thanks for a good preemptive answer :) Yes if you are good enough in identifying the "golden" cognitive processes. While it is clear you would be better than random chance, it is very unclear to me how good you would be. *

I think its worth digging into an example in detail: if you look a at early Einstein, you actually see someone with an unusually developed geometric thinking and the very lucky heuristic of interpreting what the equations say as the actual reality. Famously special relativity transformations were written first by Poincare. "All" what needed to be done was to take it seriously. General relativity is a different story, but at that point Einstein was already famous and possibly one of the few brave enough to attack the problem.

Continuing with the same example, I would be extremely doubtful if Einstein would be picked by selection process similar to what CEA or 80k hours will be probably running, before he become famous. 2nd grade patent clerk? Unimpressive. Well connected? No. Unusual geometric imagination? I'm not aware of any LessWrong sequence which would lead to picking this as that important :) Lucky heuristic? Pure gold, in hindsight.

(*) At the end you can take this as an optimization problem depending how good your superior-cognitive-process selection ability is. Let's have a practical example: You have 1000 applicants. If your selection ability is great enough, you should take 20 for individual support. But maybe its just good, and than you may get better expected utility if you are able to reach 100 potentially great people in workshops. Maybe you are much better than chance, but not really good... than, maybe you should create online course taking in 400 participants.

---

*Ben (me):*

Examples are totally worth digging into! Yeah, I actually find myself surprised and slightly confused by the situation with Einstein, and do make the active predictions that he had *some*strong connections in physics (e.g. at some point had a really great physics teacher who'd done some research). In general I think Ramanujan-like stories of geniuses appearing from nowhere are not the typical example of great thinkers / people who significantly change the world. If I'm I right I should be able to tell such stories about the others, and in general I do think that great people tend to get networked together, and that the thinking patterns of the greatest people are noticed by other good people before they do their seminal work cf. Bell Labs (Shannon/Feynman/Turing etc), Paypal Mafia (Thiel/Musk/Hoffman/Nosek etc), SL4 (Hanson/Bostrom/Yudkowsky/Legg etc), and maybe the Republic of Letters during the enlightenment? But I do want to spend more time digging into some of those.

To approach from the other end, what heuristics might I use to find people who in the future will create massive amounts of value that others miss? One example heuristic that Y Combinator uses to determine who in advance is likely to find novel, deep mines of value that others have missed is whether the individuals regularly build things to fix problems in their life (e.g. Zuckerberg built lots of simple online tools to help his fellow students study while at college).

Some heuristics I use to tell whether I think people are good at figuring out what's true, and make plans for it, include:

- Does the person, in conversation, regularly take long silent pauses to organise their thoughts, find good analogies, analyse your argument, etc? Many people I talk to take silence as a significant cost, due to social awkwardness, and do not make the trade-off toward figuring out what's true. I always trust the people more that I talk to who make these small trade-offs toward truth versus social cost
- Does the person have a history of executing long-term plans that weren't incentivised by their local environment? Did they decide a personal-project (not, like, getting a degree) was worth putting 2 years into, and then put 2 years into it?
- When I ask about a non-standard belief they have, can they give me a straightforward model with a few variables and simple relations, that they use to understand the topic we're discussing? In general, how transparent are their

models to themselves, and are the models general simple and backed by lots of little pieces of concrete evidence?
- Are they good at finding genuine insights in the thinking of people who they believe are totally wrong?

My general thought is that there isn't actually a lot of optimisation process put into this, especially in areas that don't have institutions built around them exactly. For example academia will probably notice you if you're very skilled in one discipline and compete directly in it, but it's very hard to be noticed if you're interdisciplinary (e.g. Robin Hanson's book sitting between neuroscience and economics) or if you're not competing along even just one or two of the dimensions it optimises for (e.g. MIRI researchers don't optimise for publishing basically at all, so when they make big breakthroughs in decision theory and logical induction it doesn't get them much notice from standard academia). So even our best institutions at noticing great thinkers with genuine and valuable insights seem to fail at some of the examples that seem most important. I think there is lots of low hanging fruit I can pick up in terms of figuring out who thinks well and will be able to find and mine deep sources of value.

**Edit**: Removed Bostrom as an example at the end, because I can't figure out whether his success in academia, while nonetheless going through something of a non-standard path, is evidence for or against academia's ability to figure out whose cognitive processes are best at figuring out what's surprising+true+useful. I have the sense that he had to push against the standard incentive gradients a lot, but I might just be false and Bostrom is one of academia's success stories this generation. He doesn't look like he just rose to the top of a well-defined field though, it looks like he kept having to pick which topics were important and then find some route to publishing on them, as opposed to the other way round.

---

*Greg Lewis subsequently also responded to Jan's comment:*

I share your caution on the difficulty of 'picking high impact people well', besides the risk of over-fitting on anecdata we happen to latch on to, the past may simply prove underpowered for forward prediction: I'm not sure any system could reliably 'pick up' Einstein or Ramanujan, and I wonder how much 'thinking tools' etc. are just epiphenomena of IQ.

That said, fairly boring metrics are fairly predictive. People who do exceptionally well at school tend to do well at university, those who excel at university have a better chance of exceptional professional success, and so on and so forth. SPARC (a program aimed at extraordinarily mathematically able youth) seems a neat example. I accept none of these supply an easy model for 'talent scouting' intra-EA, but they suggest one can do much better than chance.

Optimal selectivity also depends on the size of boost you give to people, even if they are imperfectly selected. It's plausible this relationship could be convex over the 'one-to-one mentoring to webpage' range, and so you might have to gamble on something intensive even in expectation of you failing to identify most or nearly all of the potentially great people.

(Aside: Although tricky to put human ability on a cardinal scale, normal-distribution properties for things like working memory suggest cognitive ability (however cashed out) isn't power law distributed. One explanation of how this could drive power-law distributions in some fields would be a Matthew effect: being marginally better than

competing scientists lets one take the majority of the great new discoveries. This may suggest more neglected areas, or those where the crucial consideration is whether/when something is discovered, rather than who discovers it (compare a malaria vaccine to an AGI), are those where the premium to really exceptional talent is less. )

---

*Jan's last response to me:*

For scientific publishing, I looked into the latest available paper[1] and apparently the data are best fitted by a model where the impact of scientific papers is predicted by Q.p, where p is "intrinsic value" of the project and Q is a parameter capturing the cognitive ability of the researcher. Notably, Q is independent of the total number of papers written by the scientist, and Q and p are also independent. Translating into the language of digging for gold, the prospectors differ in their speed and ability to extract gold from the deposits (Q). The gold in the deposits actually is randomly distributed. To extract exceptional value, you have to have both high Q and be very lucky. What is encouraging in selecting the talent is the Q seems relatively stable in the career and can be usefully estimated after ~20 publications. I would guess you can predict even with less data, but the correct "formula" would be trying to disentangle interestingness of the problems the person is working on from the interestingness of the results.

(As a side note, I was wrong in guessing this is strongly field-dependent, as the model seems stable across several disciplines, time periods, and many other parameters.)

Interesting heuristics about people :)

I agree the problem is somewhat different in areas not that established/institutionalized where you don't have clear dimensions of competition, or the well measurable dimensions are not that well aligned with what is important. Loooks like another understudied area.

[1] Quantifying the evolution of individual scientific impact, Sinatra et.al. Science, http://www.sciencesuccess.org/uploads/1/5/5/4/15543620/science_quantifying_aaf5239_sinatra.pd

# Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

This is a linkpost for https://arxiv.org/abs/1712.01815

# The World as Phenomena

This is a linkpost for https://mapandterritory.org/the-world-as-phenomena-47d27d593016

*NB: Originally posted to Map and Territory on Medium, so some of the internal series links go there.*

*NB: I made some mistakes in this post about my use of academic philosophy terms in ways that I would no longer endorse. I've let it stand for historical purposes, but of particular import is to note that "existential phenomenology" fails to capture all of my perspective well, and my understanding of idealism in this post is rather poor, since my perspective is well described, in part, by idealism, specifically phenomenalism.*

My introduction to axiological alignment is just that—an introduction. Out of necessity it covers the philosophical foundations of my thinking on AI alignment, but it doesn't explore those foundations in much detail. I could just keep writing about axiological (now noematological) alignment and AI, but I suspect I would lose everyone, so first I'm going to go back and explicate the background theory as thoroughly as seems necessary. We'll begin at the beginning, employ examples and metaphors, and explore, in order, the foundations of phenomenology, phenomenological methods, the phenomenological reduction, feedback, qualia, and finally noematology.

I think of phenomenology as the philosophy of the adept beginner. By this I mean it is the philosophy of the person skilled in all the techniques of philosophy from logic to observation to intuition who nevertheless chooses to approach philosophy as if they were a beginner so that, rather than trying to explain anything, they rest on the questions until they answer themselves. Phenomenologists are not the first to try this—Socrates, Zhuangzi, Nagarjuna, Descartes, and Hegel are among our predecessors—but we have the advantage of standing on their shoulders. Let's see how far we can see!

Consider the fundamental question of philosophy, "why?".

> "Okay, why?"
>
> *"Why" what?*
>
> "Why is anything?"
>
> *What is this "anything" you refer to?*
>
> "The world. Why does the world exist?"
>
> *Does it? How do you know?*
>
> "I'm in it!"
>
> *Oh, so why are you in the world?*
>
> "I…I don't know. I just am."

It's from not knowing why "I am" that we start because if we follow any line of questioning far enough we'll eventually need to know what this "I" is to give a

complete answer. Even if any particular answer doesn't seem to depend on "I", answers are given and understood by an "I", and we can always ask of any answer "Why do I think that?", so we have to address "I" sooner or later. Of course, we also have to deal with the whole world sooner or later, so choosing to start with "I" is also a pragmatic choice because it's something we are each quite familiar with and have privileged knowledge of, and as we'll see later there is something epistemically irreducible about "I" that makes it interesting.

Now depending on your background and education, you might be tempted to dismiss this idea of starting with "I" and jump straight to taking an objective approach. I understand the appeal, but remember we are working from a beginner's perspective, and even the logic of science must begin with the subjective. If we don't actually need "I" and it evaporates as we develop our understanding, then so be it, but beginning as naifs we must include the naive conception of "I" for now and only remove it later if we can fully account for it.

Continuing our line of questioning,

> *How do you know you are? Don't think too hard; just give the obvious answer.*
>
> "I see and hear. I think. I remember. I experience."

If we take up the kind of the radical philosophical skepticism introduced by the Ajnana school and Pyrrho, one of the few things we can claim to know is that we experience. It may not be clear who we are, what we experience, or how we experience it, but to know is to put the "I" in relation to the world and we call that relationship "experience". Existential phenomenology is the philosophy that develops if you choose to assume experience is the only source of knowledge.

There are other options. Platonism, for example, chooses differently and permits the existence of direct knowledge without the need for it to be experienced. Some strains of Gnostic and Buddhist thought suppose there is only direct knowledge and experience is an illusion. And solipsism rejects nearly everything we think of as knowledge, granting only the existence of the "I". All of these are possible ways to address "why am I?", so why phenomenology?

Two reasons. First, modern physics makes it abundantly clear that there is probably no direct knowledge. We get no faster than light travel, no true spooky action at a distance, no nothing! We always pay for our lunch, even when it's free, and if there is direct knowledge, then we've gone a suspiciously far way to understanding the physics of our universe without discovering it. At the same time, what physics does show us is that information only exists when it's moving, which is to say when measurement happens, and thus appears tightly bound to the mechanisms of causality. So taken in whole, physics paints the picture of a world where the only way to know anything is through experience.

Second, parsimony. Reasonable people may disagree, but my take is that the phenomenological perspective—that experience is the only source of knowledge—is the least complex solution that is able to fully address the question of how we know of our own existence. If we try to get away with less complexity and say experience is not necessary, I think we fail to adequately explain why it appears to us that we have a shared, objective existence that extends beyond our own knowledge. And if we try to demand more complexity by, for example, assuming the existence of direct knowledge, I think we get a gear that doesn't turn the machine and so can leave it out of our understanding without losing anything. Thus, by granting experience sole

ownership of knowledge generation, we produce a “just enough” explanation that neither has anything missing nor anything extra.

I’ll have more to say on parsimony when we look closer at the phenomenological reduction, so for now let’s wrap up our dialogue.

> *So you experience, but what if you take away the “you”? Can there be unattached experience?*
>
> “…no? …no. No, there can’t.”

Experience is always experience of something, by something. That is, it is always that some subject experiences some object. In this way we say that experience is intentional because experience is directed and exists in tension between the subject and object the same way a rope may be held in tension between two posts. If we take away the subject or the object the experience falls away the same way the rope would go slack if we took away one of the posts, and if we take away the experience then there’s nothing connecting the subject and object so they would just be things, not subjects or objects, just as our two posts would be disconnected without the rope stretched between them. Taken together we call this relation of subject, experience, and object a phenomenon.

Suppose we take a specific phenomenon, like “dog barks at car”, and try to show it is not intentional by separating out its parts. Let’s start with the subject. If we take out the dog we have only “barks at car”, but even without the dog we must imagine something is doing the barking. This gives us with a pattern that can be matched by a subject, but then we are still describing the experience of something barking at a car, so we still have a complete phenomenon, albeit one with a thought as the subject rather than a thing. As we’ll see when we address qualia, the subject being a thought poses no problem because the subject, when experienced as the subject of a phenomenon, is always a thought anyway.

If we can’t fully remove the subject, what about the object? Taking out the car nominally gives us “dog barks” and, at least in English, this is a valid construction where the experience/verb has no object. But is there really no object? Consider what barking is: the dog’s bark happens when the dog uses its body to create compression waves in the air that we identify as barking, so “barks” has an implicit object like the air or the world into which the dog does its barking. Thus we can’t really remove the object either, although we can change our frame of reference to see the object differently.

Since we can’t remove the subject and we can’t remove the object, we obviously can’t remove both and get a pure experience of barking, but what if we remove the experience and attempt to hold the subject and object as things-in-themselves? This almost seems possible, but ask yourself how you know about the dog and the car. Since we assumed that experience is the only source of knowledge, it must be that something experiences the dog and car to know about them, and in this case that thing is you! Maybe the dog and car can exist without being known, but in existential phenomenology their existence is fundamentally unknowable if they are not

experienced by a subject. So even if the dog is not related to the car by the dog's barking, the dog and the car are objects of your own experience and thus part of some phenomena. If they were not they would be literally unknown.

Thus we see that phenomena, although they have parts, are not divisible, and all knowledge exists as phenomena. This is the keystone of the phenomenological perspective.

*I see. You were secretly a phenomenologist all along!*

Don't be too surprised if you find yourself nodding along and thinking this all sounds obvious. It is! The trouble comes with the philosophical bullets you'll be asked to bite when taking this radically naive view.

Consider the question of whether a thing exists in its own right independent of phenomena. That is, do things exist if they are not experienced? For example, suppose there is a star so far away that the light from it will never reach us. Since it is outside our light cone we will never experience evidence of it, so in what sense can this hypothetical star be said to exist? From the phenomenological perspective, we might say that the *idea* of such a star exists but that the star itself does not since we can't interact with it.

Or consider particle physics. In what sense are quarks or strings or anything else "real"? Scientific realism holds that physical models, to the extent they are accurate, describe things as they really are—viz. if atomic theory correctly describes the world then the world really is made up of actual atoms. But this is to suppose a direct knowledge of the world which we can uncover through scientific inquiry. A phenomenologist might instead say we can investigate the ontic being of a thing, but only because we have knowledge of the ontological and through that may infer something of the metaphysical. This leaves phenomenology compatible with physicalism, but finds it decidedly opposed to realism and, in its existential form, idealism.

If none of that sounded too weird, when we get around to discussing qualia and consciousness we'll find that existential phenomenology implies something like functionalism and definitely implies panpsychism. I won't try to convince you of that now, but know that our humble choice to accept that all knowledge comes from experience will lead us to places far off the beaten path. It should be an exciting ride!

# Thinking as the Crow Flies: Part 1 - Introduction

# Preamble

I've wanted to write a series of posts here on logic and the foundations of mathematics for a while now. There's been some recent discussion about the ontology of numbers and the existence of mathematical entities, so this seems as good a time as any to start.

Many of the discussed philosophical problems, as far as I can tell, stem from the assumption of formalism. That is, many people seem to think that mathematics is, at some level, a formal logic, or at least that the activity of mathematics has to be founded on some formal logic, especially a classical one. Beyond that being an untenable position since Gödel's Incompleteness Theorems, it also doesn't make a whole lot of intuitive sense since mathematics was clearly done before the invention of formal logic. By abandoning this assumption, and taking a more constructivist approach, we get a much clearer view of mathematics and logic as a whole.

This first post is mostly informal philosophizing, attempting to describe exactly what logic and mathematics is about. My second post will be a more technical discussion accounting for the basic notions of logic.

# Intuitions and Sensations

To begin, I'd like to point out a fact which most would find obvious but has, in the past, lead to difficult philosophical problems. It is clear that we don't have direct access to the real world. Instead, we have senses which feed information, even if dishonestly, to our mind. These senses may be predictable, may be potentially modeled by a pattern which mimics our stream of senses. At some level, we have direct access to a sensory signal. This signal is not a pure, unfiltered lens on the world, but it is a signal, independent of, but directly accessible by, us.

We also have access to our intuitions, the part of our thoughts which we may label "ideas". We may not have total access to all our faculties. Much of our mental processing is done outside the view of our awareness. If I asked you to name a random city, eventually you'd come up with one. You'd, however, be hard-pressed to produce a trace of how that city's name came to your awareness. Perhaps you could offer background information, explaining why you'd name that city, among all the possibilities. Regardless of such accounts, we'd still lack a trace of the signal sent by your consciousness ("I want the name of a city, any city") reaching into the part of your mind capable of fulfilling such a request, and the subsequent reception of the name into your awareness. We don't have such detailed access to the inner-workings of our mind.

It seems that those things which we have direct access to are, in fact, part of us. Those intuitions within our awareness, those filtered signals which we directly experience, make up our qualia, are instantiated in the substrate of our

consciousness. They may be thought of as part of ourselves, and to say we have access to them is to say that we have direct access to those parts of ourselves of which we are aware. This, I think, is trivially true. Though that isn't essential for the rest of this piece.

We may distinguish normal intuitions from senses by the degree we can control them. Intuitions are controllable and manipulable by ourselves, while senses are not. This isn't perfectly clean. One may, for example through small DMT doses, cause one to experience controllable hallucinations which are a manifestation of direct (though not complete) control of the senses. Also, there are plenty of examples of intuitions which we find difficult to control, such as ear-worms. For the sake of this work, I will ignore such cases. What I want to focus on are sensory sensations fed to our awareness passively and those intuitions which we have complete (or complete for practical purposes) control. These are the sorts of things needed for logic, mathematics, and science, which will be the primary focuses of this series. For the remainder, by "sense" and "sensory data" I am referring to those qualia which are experienced passively, without deliberate control; by "intuition" I am referring to those intuitions which are under our direct and (at least apparent) total control.

# Grounding

At this stage, it's useful to make a remark about language and grounding. Consider what I might be saying if I describe something as an elephant. Within my mind is an intuition which I'm assigning the word "elephant", and in calling something presumed external to me an elephant, I am asserting that my intuition is an approximate model for the thing I'm naming. The difference between the intuition and the real thing is important. It is practically impossible to have a perfect understanding of real-world entities. My intuition tied to "elephant" does not contain all that I might consider knowable about elephants, but only those things which I do know. A veterinarian specializing in elephants would certainly have a more accurate, more elaborate intuition assigned to "elephant" than a non-specialist, and this wouldn't be the full extent to which elephants could be modeled. In essence, I'm using this modeling intuition as a metaphor for an elephant whenever I use that word.

Based on this, we can account for learning and disagreement. Learning can be characterized as the process of refining an approximately correct intuition modeling something external. A disagreement stems from two main places. Firstly, two people with similar sensations may be using differing models. From these differences, two people may describe identical sensations differently, as their models might disagree. Secondly, two people may think they're getting similar sensations when they are not, and so disagree because they are unable to correctly compare models, to begin with. This is the "Blind men and an elephant" scenario.

This account also cleanly explains why we can still meaningfully talk about elephants when none are present. In that case, we are speaking of the intuition assigned to "elephant". Additionally, we can talk about non-existent entities like unicorns unproblematically, as such things would still have realities as intuitions. An assertion of existence or nonexistence is really about an intuition, a model of something. The property of existence corresponds to a prediction of presence in the real world by our model, non-existence to our model predicting absence. The correctness of these properties is precisely the degree to which they accurately predict sensory data.

Intuitions need not be designed to model something in order for it to be used to model something else. If I try to describe an animal which I'm only the first time encountering, then I may construct a new model of it by piecing together parts of older models. I may even call it an "elephant-like-thing" if I feel it has some, if limited, predictive power. In this way, I'm constructing a new model by characterizing the degree to which other models predict properties of the new animal I'm seeing. Eventually, I may assign this new model a word, or borrow a word from someone else.

One can also create intuitions without attempting to model something external. If you were a mind in a void, without any sensory information, you should still be able to think of basic mathematical and logical concepts, such as numbers. You might not be motivated to do so, but the *ability* to do so is what's relevant here. These concepts can be understood in totality as intuitions, completely definable without external referents. Later, this will be elaborated on at length, but take this paragraph as-is for the moment.

Even if an intuition was created without intent to model, it still can be used as such. For example, one can think of "2" without using it to model anything. One can still say that a herd of elephants has 2 members, using the intuition of 2 as a metaphor for some aspect of the herd.

Some notion I've heard before is that it seems like a herd with 2 members would have 2 members even if there was no one around to think so, and so 2 has to exist independently of a mind. Under my account, this statement fails to understand perspective. It is certainly the case that one could model a herd of 2 using 2, regardless of if anyone else was thinking of the herd. However, even asking about the herd presupposes that at least the asker is thinking about the herd, disproving the premise that the herd isn't being thought about. If it were truly the case that no one was thinking of it at all, then there's nothing to talk about. The question would not have been asked in the first place, and the apparent problem then vanishes. It is clear at this point that stating "a herd has 2 members" does not make 2 part of our model of the world.

At this point, I will introduce terminology which distinguishes between the two kinds of intuitions discussed. Intuitions which are potentially incomplete, designed to model external entities will be called *grounded* intuitions. Those intuitions which may be complete and may exist without modeling properties will simply be ungrounded intuitions.

One common description of reality stemming from Platonism is that of an imperfect shadow or reflection of the transcendental world of ideals. After all, circles are perfect, but nothing in the world described as a circle is a truly perfect circle. By my account, perfection doesn't come into the picture. A circle is an ungrounded intuition. An external entity is only accurately called a circle in so far as the intuition of a circle accurately models the entity's physical form. The entity isn't imperfect, in some objective sense. Rather, the grounded intuition of that entity is simply more complex than the ungrounded intuition of the circle. The apparent imperfection of the world is only a manifestation of its complexity. Grounded intuitions tend to be more complicated than the ungrounded intuitions which we used to approximate the real world. This is, at once, not surprising, but significant. If we lived in an extremely simple world (or one which was simple relative to our minds) then we might create ungrounded intuitions which were simpler than the average ungrounded one. We may then have trouble distinguishing between sensory data and intuition, as all facts about the real world would be completely obvious and intuitively predictable.

# Ontological Commitments of Ungrounded Entities

I think it's worth taking the time to discuss some content related to ontological commitments and conventions. Ontological commitments were introduced by Quine, but I won't hold true to the notion as he originally described it. Instead, by an ontological commitment, I am referring to an assertion of the objective existence of an entity which is independent of the subjective experience of the person making the assertion.

Let's take a scenario where two people are arguing over what color the blood of a unicorn is. One says silver, the other red. Our goal is to make sense of this argument. Assuming neither people believe unicorns exist, what content does this argument actually have?

First, it behooves us to make sense of what a unicorn is, and what commitments we make in talking about them. For the moment, I'll stick to a conventional distributive-semantical characterization of meaning (I plan on making a post about this quite some time from now). Through our experience, we eventually associate words like "blood", "horse", and "horn" with vectors inside of some semantic space. We can then combine them in a sensical way to produce the idea of a horse with a horn, a new vector for a new idea, a unicorn. When talking about commitments, we need to make a distinction between two things; commitments to expectations, and commitments to ideas. When we define unicorns in this manner, we are committing ourselves to the idea of unicorns as something that's coherent and legible. We are not making a commitment to unicorns existing for real, that is we do not suddenly expect to see a unicorn in real life. This may be considered an ontological commitment of a sort. We certainly ascribe existence to the *idea* of a unicorn, at least within our own mind. We don't, however, ontologically commit ourselves to what the idea of unicorns might theoretically model. Since all sentences cannot help but refer to ideas rather than actual entities, regardless of our expectations, the assertion that unicorn blood is silver pertains to this idea of unicorns, nothing that exists outside of our mind.

I'd like to digress momentarily to talk about this standard conundrum:

> If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has a standard solution that I'd consider universally satisfactory. Ultimately, the question isn't about reality, it's about the definition of the word "sound". If by "sound" the asker is speaking of a sensation in the ear, then the answer is "no". If they mean vibrations in the air, then the answer is "yes". Under the distributional semantics of the word "sound", we can talk about this word having values in various directions. For some people, "sound" is assigned the region defined by a positive value in the direction corresponding to sensations in the ear. For others, "sound" is assigned to the region with positive value in the direction corresponding to vibrations in the air. These two regions have heavy overlap in practice. When we experience a sensation, it's rare for it to have a positive value in one of these, but not the other. And so, we assign one of these regions the word "sound", most of the time having no problem with others who make a different choice but arriving at disagreements over questions like the above.

But which is it? What does "sound" actually mean? Well, that's a choice. Consider the situation in detail. Is there anything that needs to be clarified? Are there vibrations in the air? Yes. Are there any sensations in an ear caused by these vibrations? No. So there's nothing left to learn. All that's left is to decide how to describe reality. It may even be useful to split the term, to talk about "type-1 sound" and "type-2 sound", which usually coincide, but don't on rare occasions. Regardless, it's a matter of convention, not a matter of fact, whether the word "sound" should apply.

And so, we're in sight of the resolution to the unicorn blood argument. One person has a region in their semantic space corresponding to one-horned horses with silver blood, and want's to assign that region the word "unicorn". The other person has identified a close-by semantic region, but there the blood is red, and they want that to have the word "unicorn". Note that neither would think that the others claim is nonsense. The argument is not predicated on, for example, one person thinking the idea of a unicorn with red blood is incoherent. Both parties agree that each other have identified meaningful regions of semantic space. They are making identical ontological commitments. What they are disagreeing on is a naming convention.

Throughout this series, I will often discuss mathematics and logic as fundamentally subjective activities, but this does not mean I reject mathematical objectivism as such. Rather, the objective character of mathematics moves from being an aspect of mathematics itself to being an aspect of how it's practiced. Mathematics is done as a social activity carried by a convention which is itself objective: or at least (ideally) as objective as a ruler. Showing that someone is mathematically wrong largely boils down to showing which convention a person is breaking in making an incorrect judgment.

Brouwer, who was the first to really push mathematical intuitionism, described mathematics as a social activity at its core. As a consequence, he argued against the idea of a formal logical foundation before Gödel's incompleteness theorems were even discovered.

The basic idea of constructivism is to limit our ontological commitments as much as possible. Consider the well known "I think, therefore I am". It highlights the fact that the act of thinking and introspection itself implies an ontological commitment to the self. Since we are already doing those things, it's really not much of a commitment at all. Similarly, the fact that I am writing in a language commits me ontologically to the existence of the language I'm writing in. As I'm doing this anyway, it's not much of a commitment. For this, I call these sorts of commitments "cheap commitments".

Mathematical and logical entities are ideas. By discussing them, we are committing ourselves to the existence of these entities at least as ideas. For example, if I say "there exists an even natural number", I am committing myself to the ideas of natural numbers and evenness. I'm also committing myself to the coherence or soundness of these ideas, that the statement in question is meaningful modulo the semantics of the ideas used.

I can easily make grammatical-looking sentences that seem to make some sort of expensive commitment. For example, I could say that g'glemors exist and that a h'plop is an example of a g'glemor on account of hipl'xtheth. If I said those things with any sort of seriousness I'd be committing myself to the existence of those mentioned things at least as ideas, as well as the soundness of those ideas. Being nonsense words not representing anything at all, I'd obviously be misguided in making such commitments, they certainly aren't cheap.

The point of a constructivist account is to describe mathematical and logical ideas in such a way that one is committed to their soundness in a cheap way. And here we can start to see the significance of characterizing mathematics and logic as being about ungrounded entities. In order for my commitments to those ideas to be cheap, they must be totally characterized by something that comes from within me, by something that I'm doing anyway when discussing those ideas.

# Precommitments and Judgments

We say that an idea is a cheap commitment if, in defining the notion, we summon the entity being defined, or perform the activity which we are judging to be the case. In order to do this, we need to pay attention to precommitments.

A precommitment is a prescription we make of our own behavior. It's an activity which is being done so long as those prescriptions are being followed. Precommitments are the core of structured thinking. Whenever we impose any pattern or consistency to our thinking, we are making a precommitment. By analyzing our precommitments closely, we can construct, explicitly, ideas which are cheap ontological commitments. If we are actively doing a precommitment, then we can cheaply acknowledge the existence of the idea conjured by this precommitment.

Many ungrounded intuitions arise as a form of meaning-as-usage. Some words don't have meaning beyond the precise way they are used. If you take a word like "elephant", it's meaning is contingent on external information which may change over time. A word like "and", however, isn't. As a result, we'd say "and"'s meaning fundamentally boils down to how it's used, and nothing more. Going beyond that, if we are to focus on ungrounded intuitions which are complete and comprehensible, then we are focusing precisely on those ungrounded intuitions who's definition is precisely a specification of usage, and nothing more. That specification of usage is our precommitment. Of course, usage happens outside the mind, but the rules dictating that usage aren't, and its those canonical rules of usage which I mean by "definition".

The basic elements of definitions are judgments. Judgments include things like judging that something is a proposition, or is a program, or is some other syntactic construction. Judgments also include assertions of truth, falsehood, possibility, validity, etc of some data. However, be aware that a judgment simply consists of a pattern of mental tokens which we may declare. Regardless of what preconceptions about possibility, truth, etc. one has, these should be overwritten by the completed meaning explanation in order to be understood as a purely ungrounded intuition and a cheap commitment.

When we make a judgment, we are merely asserting that we may use that pattern in our reasoning. Precommitments, as we will make use of them here, are a collection of judgments. As a consequence, what we are precommitting ourselves to is an allowance of usage for certain patterns of mental tokens when reasoning about a concept. The full precommitment summoning some concept will be called the meaning explanation for that concept.

Ultimately, it is either the case that we make a particular judgment or we don't. That, however, is a fact about our own behavior, not about the nature of reality in total, in essence. Furthermore, someone not making a particular judgment is not automatically making the opposite, or negated, judgment. In fact, such a thing doesn't even make sense in general. As a result, we don't reproduce classical logic. Though, as we'll

eventually see, there are constructive logics which are classical. However, it's worth dispelling the idea that there's "one true logic". Questions about which kind of logical symbols, classical, intuitionistic, linear, etc. is the "true" one are nonsense. One is only correct relative to some problem which has an element which is to be modeled by one of these. Whichever is the more accurate model is the correct one, there is no "one true logic", and it's certainly not the case that the intuitions which make up mathematics are governed by a classical logic. For example, the existence of theoretically unsolvable problems (e.g. the halting problem) illustrates that our capacity for judging truth is fundamentally constrained, not by some objective transcendental standard for truth, but rather by our ability to make proofs.

To summarize, to define a concept we give a list of judgments, rules dictating which patterns of tokens we can use when considering the concept. So long as these rules are being followed, the concept exists as a coherent idea. If the precommitment is violated, for example by making a judgment about the concept which is not prescribed by the rules, then the concept, as defined by the original precommitment, no longer exists. There may be a new precommitment that defines a different concept using the same tokens which is not violated, but that, being a different precommitment, constitutes a different meaning explanation, and so its summoned concept does not have the same meaning. So long as I follow a precommitment defining a concept, it is hypocritical of me to deny the coherence of that concept, just as it would be hypocritical to deny my language as I speak, to deny my existence so long as I live.

# Computation to Canonical Form

We are now free to explore an example of the construction of an ungrounded intuition. I should be specific and point out that not all ungrounded intuitions are under discussion. For the sake of mathematics and logic, intuitions must be completely comprehensible. Unlike grounded intuitions, an ungrounded one may be such that it's never modified by new information. This doesn't describe all ungrounded intuitions, but it describes the ones we're interested in.

One of the most important judgments we will consider is of the form $a \downarrow b$. It is a kind of computational judgment. It's worth explaining why computation is considered before anything else in mathematics. To digress a bit, it’s easy to argue that some notion of computation is necessary for doing even the most basic aspects of ordinary mathematics. Consider, for example, the standard theorem; for all propositions X and $Y$, $X \wedge Y = Y \wedge X$. The universal quantification allows us to perform a substitution, getting, for example, $(A \wedge C) \wedge B = B \wedge (A \wedge C)$, as an instance.

We should meditate on substitution, an essential requirement of even the most basic and ancient aspects of logic. Substitution is an algorithm, a computation which must be performed somehow. In order to realize $(A \wedge C) \wedge B = B \wedge (A \wedge C)$, we must be doing the activity corresponding to the substitution of X with $(A \wedge C)$ and the action corresponding to the substitution of $Y$ with B at some point. Substitution will appear

over and over again in various guises, acting as a central and powerful notion of computation. To emphasize, once substitution is available, we are 90% of the way toward complete and fully general Turing-Complete computation via the lambda calculus. Much of the missing features pertain to explicit variable binding, which we need anyway in order to use the quantifiers of first-order logic. I don't think it's really debatable that computation ontologically precedes logic. One can do logic as an activity, and much of that activity is computational in nature.

Before expositing on some example judgments, we should address the need for isolating concepts. Consider a theory with natural numbers N and products $A \times B$. We must ask what constitutes a natural number and a product. By default, we can form a natural number as either zero or the successor of a natural number. e.g. 0, S0, SS0, SSS0, ... A product can be formed via $(a, b)$ where a is an A and b is a B. Additionally, we have that, if a is a natural number then $\pi_1(a, b)$ (where $\pi_1$ is a projection function) is a natural number, and if b is a natural number then $\pi_2(a, b)$ is a natural number, and if b is a natural number then $\pi_1(\pi_2(a, (b, c)))$ is a natural number, etc. to infinity. This situation gets branchingly more complex as we add new concepts to our theory. If we don't define concepts as fundamentally isolated from each other, we inhibit the extensibility of our logic. This is both unpragmatic and unrealistic, as we will want to extend the breadth of concepts we can deal with as we model more novel things. Furthermore, the coherence of the concept of a natural number should not depend on the coherence of the notion of a product. Ultimately, each concept should be defined by some precommitment consisting of a list of rules for making judgments. If we entertain this infinite regress, then there may be no way in general to state what the precommitment in question even is.

At the core of our definitions will be canonical forms. Every time we define a new concept, we will assert what its canonical forms are. For example, in defining the natural numbers we will judge that $0 \in N$ and that, assuming $n \in N$, we can conclude that $S(n) \in N$. We can't assume this alone, however. Consider, for example $2 + 3$, which should be a natural number, but isn't in the correct form. We now have an opportunity to explain $\downarrow$. $a \downarrow b$ indicates that we start out with some mental instantiation a, and after some mental attention, it becomes the instantiation b. So we have, for example $1 + 1 \downarrow 2$. When I say $1 + 1 \downarrow 2$, I do not mean that $1 + 1$ is equal to 2. That's a separate kind of judgment. This means our full judgment is that $n \in N$ iff $n \downarrow 0$ or $n \downarrow S(m)$ for some $m \in N$. There are some details missing from this definition, but it should serve as a guiding example, the first rough sketch of what I mean by a meaning explanation.

It is worth digressing somewhat to critique the axiomatic method. Most people, especially when first learning of a subject, will experience a mathematical or logical

concept as a grounded intuition. This is reflected in a person's answer to questions such as "why is addition commutative?". Most people could not answer. It is not part of the definition of addition or numbers for this property to hold. Rather, this is a property stemming from more sophisticated reasoning involving mathematical induction. A person can, none the less, feel an understanding of mathematical concepts and an acceptance of properties of them without knowledge of their underlying definitions. Axiomatic methods, such as the axioms of ZFC, don't actually define what they are about. Instead, they list properties that their topic must satisfy.

The notion of ZFC-set, in some sense, is grounded by an understanding of the axioms, though it is still technically an ungrounded intuition. This state of affairs holds for any axiomatic system. There is something fundamentally ungrounded about a formal logic, but it's not the concepts which the axioms describe. Rather, what we have in a formal logic is a meaning explanation for the logic itself. That is, the axioms of the logic tell us precisely what constitutes a proof in the logic. In this way, we may formulate a meaning explanation for any formal logic, consisting of judgments for each axiom and rule of inference. Consequently, we can cheaply commit ourselves to the coherence of the logic as an idea. What we can't cheaply commit ourselves to are the ideas expressed within the logic. After all, a formal logic could be inconsistent, it's ideas may be incoherent.

As a consequence, the notion of a coherent idea of ZFC-set cannot be committed to cheaply. This holds similarly for any concept described purely in terms of axioms. It might be made cheap by appealing to a sufficient meaning explanation, but without additional effort, things treated purely axiomatically lack proper definitions in the sense used here.

# Mana

(This is a copy-pasted blog post from https://sinceriously.fyi/, I'm checking how much demand here there is for my writing.)

This is theorizing about how mana works and its implications. (Edit: usable link: https://sinceriously.fyi/wp-content/uploads/2017/12/Brent-Dill-So-wacky-RPG-session-from-a-riff-with-Anisha-and..pdf)

Some seemingly large chunks of stuff mana seems to be made of:

- Internal agreement. The thing that doles out "willpower".
- Ability to not use the dehumanizing perspective in response to a hostile social reality.

I've been witness to and a participant in a fair bit of emotional support in the last year. I seem to get a lot less from it than my friends. (One claims suddenly having a lot more ability to "look into the dark" on suddenly having reliable emotional support for the first time in a while, leading to some significant life changes.) I think high mana is why I get less use. And I think I can explain at a gears level why that is.

Emotional support seems to be about letting the receiver have a non-hostile social reality. This I concluded from my experience with it, without really having checked against common advice for it, based on what seems to happen when I do the things that people seem to call emotional support.

I googled it. If you don't have a felt sense of the mysterious thing called "emotional support" to search and know this to be true, then from some online guides, here are some supporting quotes.

From this:

- "Also, letting your partner have the space he or she needs to process feelings is a way of showing that you care."
- "Disagree with your partner in a kind and loving way. Never judge or reject your mates ideas or desires without first considering them. If you have a difference of opinion that's fine, as long as you express it with kindness."
- "Never ignore your loved one's presence. There is nothing more hurtful than being treated like you don't exist."

From this:

- "Walk to a private area."
- "Ask questions. You can ask the person about what happened or how she’s feeling. The key here is to assure her that you’re there to listen. It’s important that the person feels like you are truly interested in hearing what she has to say and that you really want to support her."
- "Part 2 Validating Emotions"
- "Reassure the person that her feelings are normal."

I think I know what "space" is. And mana directly adds to it. Something like, amount of mind to put onto a set of propositions which you believe. I think it can become easier to think through implications of what you believe is reality, and decide what to do,

when you're not also having part of you track a dissonant social reality. I've seen this happen numerous times. I've effectively "helped" someone make a decision just by sitting there and listening through their decision process.

The extent to which the presence of a differing social reality fucks up thinking is continuous. Someone gives an argument, and demands a justification from you for believing something, and it doesn't come to mind, and you know you're liable to be made to look foolish if you say "I'm not sure why I believe this, but I do, confidently, and think you must be insane and/or dishonest for doubting it", which is often correct. I believe loads of things that I forget why I believe, and could probably figure out why, often only because I'm unusually good at that. But you have to act as if you're doubting yourself or allow coordination against you on the basis that you're completely unreasonable, and your beliefs are being controlled by a legible process. And that leaks, because of buckets errors between reality and social reality at many levels throughout the mind. (Disagreeing, but not punishing the person for being wrong, is a much smaller push on the normal flow of their epistemology. Then they can at least un-miredly believe that they believe it.)

There's a "tracing the problem out and what can be done about it" thing that seems to happen in emotional support, which I suspect is about rebuilding beliefs about what's going on and how to feel about it, independent of intermingling responsibilities with defensibility. And that's why feelings need to be validated. How people should feel about things is tightly regulated by social reality, and feelings are important intermediate results in most computations people (or at least I) do.

Large mana differences allow mind-control power, for predictable reasons. That's behind the "reality-warping" thing Steve Jobs had. I once tried to apply mana to get a rental car company to hold to a thing they said earlier over the phone which my plans were counting on. And accidentally got the low-level employee I was applying mana to to offer me a 6-hour car ride in her own car. (Which I declined. I wanted to use my power to override the policy of the company in a way that did not get anyone innocent in trouble, not enslave some poor employee.)

The more you shine the light of legibility, required defensibility and justification, public scrutiny of beliefs, social reality that people's judgement might be flawed and they need to distrust themselves and have the virtue of changing their minds, the more those with low mana get their souls written into by social reality.  I have seen this done for reasons of Belief In Truth And Justice. Partially successfully. Only partially successfully because of the epistemology-destroying effects of low mana. I do not know a good solution to that. If you shine the light on deep enough levels of life-planning, as the rationality community does, you can mind control pretty deep, because almost everyone's lying about what they really want. The general defense against this is akrasia.

Unless you have way way higher mana than everyone else, your group exerts a strong push on your beliefs. Most social realities are full of important lies, especially lies about how to do the most good possible. Because that's in a memetic war-zone because almost everyone is really evil-but-really-bad-at-it. I do not know how to actually figure out much needed original things to get closer to saving the world while stuck in a viscous social reality.

I almost want to say, that if you really must save the world, "You must sever your nerve cords. The Khala is corrupted". That'll have obviously terrible consequences, which I make no claim you can make into acceptable costs, but I note that even I have

done most of the best strategic thinking in my life in the past year, largely living with a like-minded person on a boat, rather isolated. That while doing so, I started focusing on an unusual way of asking the question of what to do about the x-risk problem, that dodged a particular ill effect of relying on (even rare actual well-intentioned people's) framings.

I've heard an experienced world-save-attempter recommend having a "cover story", sort of like a day job, such as... something something PhD, in order to feel that your existence is justified to people, an answer to "what do you work on" and not have that interfering with the illegibly actually important things you're trying. Evidence it's worth sacrificing a significant chunk of your life just to shift the important stuff way from the influence of the Khala.

Almost my entire blog thus far has been about attempted mana upgrades. But recognizing I had high mana before I started using any of these techniques makes me a little less optimistic about my ability to teach. I do think my mana has increased a bunch in the course of using them and restructuring my mind accordingly, though.

# Conceptual Similarity Does Not Imply Actionable Similarity

This is another essay about naming things, dichotomies, and where subtle mix-ups can lead to errors. More specifically, I'd like to draw your attention to situations where a very real conceptual commonality is present between several problems, but this commonality doesn't actually provide much insight into a unified *solution* for the aforementioned problems.

Concretely, we can refer to time-inconsistent preferences, the well-documented phenomenon where we'll relent to in-the-moment urges, often for a temptation we will later regret. For example, a student might put off studying until the last moment, choosing instead to read a riveting novel. Or a partygoer might drink far too much they can handle, knowing they'll soon end up regretting it.

In both of these cases, there is indeed something we can abstract from the nature of each of these situations—a human considers doing X and soon regrets it, instead wishing they had done Y. My claim here is that "time-inconsistent preferences" form a type of descriptive classification because they can help us see the larger shape of what's going on, but they *don't* tell us how to solve the general problem.

Or, more specifically, I claim that in these situations where you've got a descriptive classification, it's actually the specific details (and not the ability to recognize that you're engaging in a general phenomenon) which provide the most leverage towards solving your problem.

In the above two examples, it might be that our struggling student needs to reexamine their priorities. Perhaps the regret is misplaced and actually doing poorly on the upcoming test isn't even that big of a deal. Or perhaps our student could rearrange their schedule around and study with a friend to shave off some of the aversion.

The point is, this ends up looking quite different from what our overzealous partygoer might want to do. Our partygoer may want to consider the sort of circumstances which brought them to said party in the first place; it might be the case that avoiding certain triggers means they could sidestep potential binge opportunities entirely.

The point is, there's a sort of mental misstep that can happen where being able to simply identify the generalized principle at work could give the false impression that you also know how to solve the problem. But the two are very much independent, as the generalized principle is, in these cases, a descriptive classification and not one that focuses on actions or implementation.

It might seem like I'm splitting hairs here—there's a sort of argument that the above might represent, the sort of thing where I argue that well *technically* everything is implementation-specific because at some point, you'll always need to get specific. After all, you can't actually directly act on advice to "remove triggering environmental cues" (unless you've got a weird set of billiard firearm fauna).

And I'm not trying to be the person who's trying to win by technicalities or definitions. Having the sense of the general shape of the sort of problem you're facing *can* be helpful. Knowing what sorts of general solutions tend to work can provide a helpful

template you can fill in with your environment-specific details. It can provide a good springboard for brainstorming ways forward.

EX: “Okay, I know that habit formation works best with strong, related sensory cue. What’s something related to flossing that I could use…?”

But I still do think that this sort of conceptual muddling can be pernicious, especially in the related case where two tasks *seem* similar, despite having vastly different effects. So let’s pivot a little bit to a slightly different situation: situations where your brain, when faced with an apparent conceptual similarity, *assumes* their actionable similarity, and then defaults to the easier one.

For example, Hunting for Practicality is about how, even when we try to internalize advice, it often gets cached in our brains in a way that’s akin to declarative semantic memory—we end up representing how the concepts are linked to each other and perhaps what properties they hold.

But, really, the information we should be trying to internalize should be procedural in nature—we care about *how* the advice can actually affect our actions in practice.

The default is to represent the information as a concept map; that’s a rather simple translation of the information presented to you that doesn’t require much additional effort. Trying to actively consider how your future actions will change as a result of heeding the advice in question is more involved.

And in the absence of other factors, the less effortful option wins out.

For second related example, In Defense of the Obvious is about how just receiving advice can set off our brain’s dismissal signals too quickly when it sounds like something we’ve heard a million times before. Yet, the advice is often still valuable even when it pattern matches to “boring” or “obvious”. Overriding the immediate dismissal response and doing the advice anyway is often a better response.

Once again, verifying whether or not you’ve heard such advice before ends up being an easier task than noting the dismissal, filing it away, and then looking into yourself to see if you actually are doing said obvious advice. It is also the more effortful one.

As a third example, consider a student trying to study math. They have a couple of choices: One thing they could do is read through the textbook and trace through the examples, making sure they can follow each step. Or, they could cover up the example problem and try to work through it themselves.

Looking through the textbook can provide the illusion of “exercising your math muscles”, but it’s the actual act of trying to solve problems which improve your ability to solve problems. And of course doing the real math is what’s harder, in terms of time and effort involved.

The issue with all three of these examples seem to be about the mental labels we use. Both the ineffective option and the active option can fall under the same category (e.g. “studying for math”), despite their major differences.

(The overall pattern here seems to be that of the Recognizing vs Generating distinction: roughly speaking, it’s much more difficult to put the pieces together at first, than to merely verify that the pieces fit.)

One view is that this sort of behavior is an example of self-signaling. After all, it's much easier to feel productive than to actually be productive. I also think that, on some level, your brain thinks that you *really can* get all of the same benefits by doing the easier thing. In this case, your brain is hopeful and also wrong.

There are two takeaways here:

One is the importance of putting in deliberate, mindful effort, a thesis I hope to expound on in a later post.

The second one is perhaps a more familiar variant on the (now) well-worn phrase "the map is not the territory". In this case, your ontology is not the reality. Inferences you make based on similarities may not transfer to other inferences *of a similar type*.

(Aka "Similarity-based connections are not themselves connected by similarities!")

# 12/12/2017 Update: Creating Sequences

Over the past few weeks, we've been pushing updates that improve LW 2.0's Sequence functionality, as well as cleaning up some of the less well known sequences.

Those updates are now complete and ready for general use. Highlights include:

## The Library

If you click on the "Recommended Reading" title on the front page, or on "The Library" in the main menu, you'll be taken to our new Library page. This includes sections for our Core Reading, Curated Sequences, and Community Sequences.

**Core Reading** is essential content that has stood the tests of time, that users are generally expected to have read.

**Curated Sequences** are additional high quality essays that the admins have decided to feature.

**Community Sequences** are where most new sequences will appear. (Currently sorted by "newest first", although we'll probably update the sorting mechanisms soon)

## Creating New Sequences

Under the "Community Sequences" header, you'll see a "Create new sequence" button. This is deliberately somewhat hard-to-find, since we don't want new users to immediately start creating sequences willy nilly.

To create a sequence, you'll start by giving it a name, description, and images for the Banner and Thumbnail. It's recommend that the Banner and Thumbnail be the same picture (although perhaps cropped differently so that they look good at large and small scales).

Banner images must be at least 1600 pixels wide. Eventually we'll add some tools to help users find appropriate images. Meanwhile, if you need help finding an appropriate banner, you can ping us and we'll help you out. (In general, a good tip is to go to Google Image search, click "tools" tab, and restrict results to images whose *size* is at least 2 MP, and whose *usage rights* are set to "noncommercial re-use)

After clicking the "create" button, you'll be able to add and change the ordering of posts.

## User Profile Sequences

Once you've created at least one sequence, your User Profile page will gain a Sequences section. This will list your sequences, and includes another "create sequence" button.

You can see an example of this at Luke Muehlhauser's user page.

# Happy Sequencing!

Creating better sequence tools is a core element of the LW 2.0 strategy. Some of our next goals are to improve your ability to keep track of which sequences you're currently reading (and for longer multi-part sequences, automatically suggesting the next sequence of a book when you finish)

Have fun creating new content!

# The expected value of the long-term future

I wrote an article describing a simple model of the long-term future. Here it is:

- in PDF format
- in tex format

Summary:

A number of ambitious arguments have recently been proposed about the moral importance of the long-term future of humanity, on the scale of millions and billions of years. Several people have advanced arguments for a cluster of related views. Authors have variously claimed that shaping the trajectory along which our descendants develop over the very long run (Beckstead, 2013), or reducing extinction risk, or minimising existential risk (Bostrom, 2002), or reducing risks of severe suffering in the long-term future (Althaus and Gloor, 2016) are of huge or overwhelming importance. In this paper, I develop a simple model of the value of the long-term future, from a totalist, consequentialist, and welfarist (but not necessarily utilitarian) point of view. I show how the various claims can be expressed within the model, clarifying under which conditions the long-term becomes overwhelmingly important, and drawing tentative policy implications.

# Philosophy of Numbers (part 1)

This post is the first in a series of things that I think would be fun to discuss on LW. Part two is here.

---

It seems like there are (at least) two kinds of things we make statements about: physical things, like apples or cities, and logical things, like numbers or logical relations. And it's pretty interesting to question how accurate this seeming is. Are numbers really a "kind of thing," and what do we mean by that anyways? Can we unify these multiple kinds of things, or kinds of statements, into one kind, or not?

For a light review of standard answers, see this nice video. For more depth, you might see the SEP on abstract objects or philosophy of mathematics.

Compare the statements "There exists a city larger than Paris" versus "There exists a number greater than 17." It seems like we use much the same thought patterns to evaluate both these statements, and both seem to be true in the same ordinary sense. Yet the statement about cities seems true because of a correspondence to the external world, but there is no "17" object in a parsimonious predictive model of the world.

To this you might say, "What's the big deal? Even if I don't think numbers are physical objects, it's perfectly reasonable to make this tight analogy between cities and numbers in our reasoning. How is making a big issue out of this going to help us do anything practical?"

Well, in logical decision theory, a recent formulation of some ideas from TDT/UDT, the agent wants to make a causal model of the world that includes (in the model) "causal" effects of a fixed mathematical statement (speficially, the output of the agent's own algorithm). First of all, this is pretty novel and we don't really know how to formalize learning such a model. Second, it's pretty philosophically weird - how is a piece of math supposed to have something like a causal effect on trees and rocks? If we want to solve the practical problem, it might help to be less confused about numbers.

Plus, you know, it's interesting! Why do we think there's such a thing as "numbers," how come the same reasoning works for both numbers and cities, and what are the limits to this analogy, if any?

---

When one wants to outdo an entire branch of philosophy, it's nice to have some sort of advantage. And the sign of such an advantage is often a bunch of philosphers being loudly wrong about some related issue. But this case, I don't see the signs of an easy advantage. Modern philosophy of numbers doesn't seem to have a bunch of sharp divides or false confidence. Instead, most everyone seems pretty aware that they're confused, despite some fairly interesting ideas being available.

But, okay, I do have some ideas.

See, if you ask philosophers about something that might exist, their first instinct is to try to find a necessary-and-sufficient definition of this thing, more or less on its own terms. Over here, we're much more likely to think of how things are represented in

people's models of the world, and ask what chain of events led people to have that representation, which I think is some important philosophical technology.

This wouldn't be a proper post without a pile of links. So here are some options we might want to keep in mind: Taboo your words. Focus on origin or function, like in the example of "truth." Imagine what cognitive algorithm you're using. Keep your eye on the reductionist ball.

---

Since this post is labeled "part 1," you might expect that I'm going to end this without telling you exactly what I think about numbers. You'd be right!

But I do want to prompt you with some questions I think are more key than "what is math, really?", and corresponding things I think might be hints.

- **Why do we say that numbers "exist?"**

  Why do we need a property called "existence" in the first place, even just for trees and rocks? Those Eliezer-posts about truth may hint at one point of view.

- **Why would we want to say that certain abstract sentences are "true?"**

  Do statements about math have the same properties Eliezer outlined as making "true" a useful word? Why would we want talk about labels on mathematical sentences if math is just a bunch of tautologies?

- **Does it make sense to evaluate "There exists a city larger than Paris" and "There exists a number greater than 17" the same way?**

  What cognitive algorithms could we be using? What are their disadvantages?

- **Does this line of reasoning actually help us implement LDT?**

  I got nothing.

# Bayes and Paradigm Shifts - or being wrong af

So I've been thinking about Bayesian probability and paradigm shifts. One of the early examples that Price published after discovering Bayes' theorem (after Bayes died) was of someone who, upon awakening for the first time with no other information on cosmology, if they knew Bayes theorem, could then update their probability that the sun would rise again the next day, each day they saw it rise again. So with time, as they see the sun rise more and more times, they become more and more 'certain' that it will rise again the next day (ie their priors become higher).

However, not having any knowledge of the universe or physics, they are unaware that there is a near certainty that this sun will someday supernova and no longer rise again. If they made thousands of generations of sun tracking bayesians, every day they would see the sun rise and update their probability, and become more certain that it would rise again. By the time it didn't rise, they would be wildly certain that it would rise again. So the more certain they became, actually the more WRONG they became. That sun was always almost certainly doomed at the same 99.999....% level the whole time (maybe not to each given new day, but eventually) and they just didn't have access to good enough priors to recognize this.

So as a result of bad priors, they are maybe increasing their accuracy relative to any given day (the sun only dies on 1 in a billion days) but decreasing their accuracy of it's eventual transformation into a black hole or some such phenomena which will likely kill the shit out of them.

I think this kind of misinformed search for accuracy is very symbolic of a bayesian look at paradigm shifts (even as it could also b used as a limited critique of bayesian statistics). Once they get access to just the knowledge that other stars exist, it opens up a huge range of other variables they didn't know about in the calculation of their priors. So while we're chugging along in our search for accuracy, we may be building relative accuracy, while building absolute an error until our paradigm catches up with a new and deeper layer of information.