Best of LessWrong: November 2019

1. Book Review: Design Principles of Biological Circuits
2. Chris Olah’s views on AGI safety
3. Evolution of Modularity
4. Gears-Level Models are Capital Investments
5. A mechanistic model of meditation
6. Mental Mountains
7. AlphaStar: Impressive for RL progress, not for AGI progress
8. The Curse Of The Counterfactual
9. Instant stone (just add water!)
10. Could someone please start a bright home lighting company?
11. The LessWrong 2018 Review
12. Can you eliminate memetic scarcity, instead of fighting?
13. But exactly how complex and fragile?
14. The Credit Assignment Problem
15. Total horse takeover
16. Will transparency help catch deception? Perhaps not
17. Antimemes
18. Explaining why false ideas spread is more fun than why true ones do
19. Robin Hanson on the futurist focus on AI
20. Relevance Norms; Or, Gricean Implicature Queers the Decoupling/Contextualizing Binary
21. Matthew Walker's "Why We Sleep" Is Riddled with Scientific and Factual Errors
22. The new dot com bubble is here: it’s called online advertising
23. AI Alignment Research Overview (by Jacob Steinhardt)
24. Wrinkles
25. Pieces of time
26. [AN #75]: Solving Atari and Go with learned game models, and thoughts from a MIRI employee
27. Neural Annealing: Toward a Neural Theory of Everything (crosspost)
28. Historical forecasting: Are there ways I can get lots of data, but only up to a certain date?
29. Epistemic Spot Check: Unconditional Parenting
30. A test for symbol grounding methods: true zero-sum games
31. Platonic rewards, reward features, and rewards as information
32. The Bus Ticket Theory of Genius
33. Thoughts on Robin Hanson's AI Impacts interview
34. Hard Problems in Cryptocurrency: Five Years Later - Buterin
35. [AN #74]: Separating beneficial AI into competence, alignment, and coping with impacts
36. How I do research
37. Autism And Intelligence: Much More Than You Wanted To Know
38. Is daily caffeine consumption beneficial to productivity?
39. Pricing externalities is not necessarily economically efficient
40. Politics is work and work needs breaks
41. Self-Keeping Secrets
42. A Practical Theory of Memory Reconsolidation
43. Units of Action
44. Personal quality experimentation
45. My Anki patterns
46. 3 Cultural Infrastructure Ideas from MAPLE
47. Operationalizing Newcomb's Problem
48. Building Intuitions On Non-Empirical Arguments In Science
49. How common is it for one entity to have a 3+ year technological lead on its nearest competitor?
50. [1911.08265] Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model | Arxiv

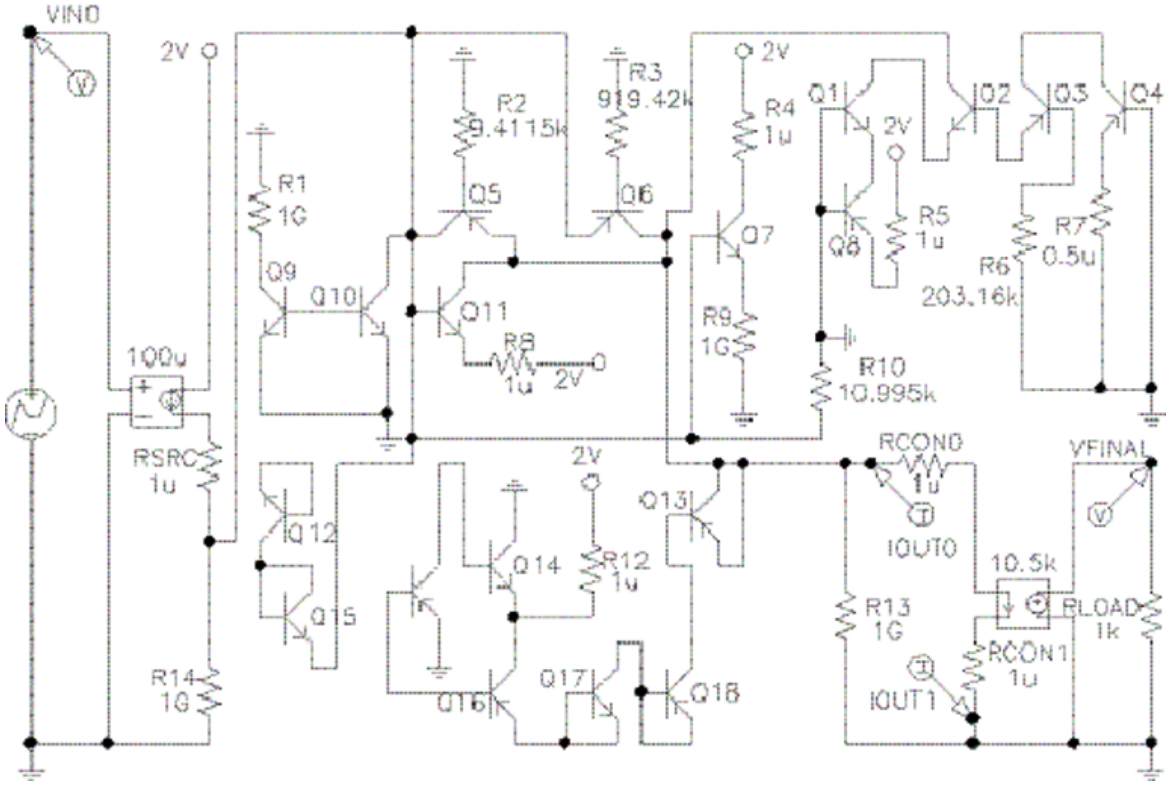
Best of LessWrong: November 2019

1. Book Review: Design Principles of Biological Circuits
2. Chris Olah’s views on AGI safety
3. Evolution of Modularity
4. Gears-Level Models are Capital Investments
5. A mechanistic model of meditation
6. Mental Mountains
7. AlphaStar: Impressive for RL progress, not for AGI progress
8. The Curse Of The Counterfactual
9. Instant stone (just add water!)
10. Could someone please start a bright home lighting company?
11. The LessWrong 2018 Review
12. Can you eliminate memetic scarcity, instead of fighting?
13. But exactly how complex and fragile?
14. The Credit Assignment Problem
15. Total horse takeover
16. Will transparency help catch deception? Perhaps not
17. Antimemes
18. Explaining why false ideas spread is more fun than why true ones do
19. Robin Hanson on the futurist focus on AI
20. Relevance Norms; Or, Gricean Implicature Queers the Decoupling/Contextualizing Binary
21. Matthew Walker's "Why We Sleep" Is Riddled with Scientific and Factual Errors
22. The new dot com bubble is here: it’s called online advertising
23. AI Alignment Research Overview (by Jacob Steinhardt)
24. Wrinkles
25. Pieces of time
26. [AN #75]: Solving Atari and Go with learned game models, and thoughts from a MIRI employee
27. Neural Annealing: Toward a Neural Theory of Everything (crosspost)
28. Historical forecasting: Are there ways I can get lots of data, but only up to a certain date?
29. Epistemic Spot Check: Unconditional Parenting
30. A test for symbol grounding methods: true zero-sum games
31. Platonic rewards, reward features, and rewards as information
32. The Bus Ticket Theory of Genius
33. Thoughts on Robin Hanson's AI Impacts interview
34. Hard Problems in Cryptocurrency: Five Years Later - Buterin
35. [AN #74]: Separating beneficial AI into competence, alignment, and coping with impacts
36. How I do research
37. Autism And Intelligence: Much More Than You Wanted To Know
38. Is daily caffeine consumption beneficial to productivity?
39. Pricing externalities is not necessarily economically efficient
40. Politics is work and work needs breaks
41. Self-Keeping Secrets
42. A Practical Theory of Memory Reconsolidation
43. Units of Action
44. Personal quality experimentation
45. My Anki patterns

46. 3 Cultural Infrastructure Ideas from MAPLE
47. Operationalizing Newcomb's Problem
48. Building Intuitions On Non-Empirical Arguments In Science
49. How common is it for one entity to have a 3+ year technological lead on its nearest competitor?
50. [1911.08265] Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model | Arxiv

Book Review: Design Principles of Biological Circuits

I remember seeing a talk by a synthetic biologist, almost a decade ago. The biologist used a genetic algorithm to evolve an electronic circuit, something like this:



(source)

He then printed out the evolved circuit, brought it to his colleague in the electrical engineering department, and asked the engineer to analyze the circuit and figure out what it did.

“I refuse to analyze this circuit,” the colleague replied, “because it was not designed to be understandable by humans.” He has a point - that circuit is a big, opaque mess.

This, the biologist argued, is the root problem of biology: evolution builds things from random mutation, connecting things up without rhyme or reason, into one giant spaghetti tower. We can take it apart and look at all the pieces, we can simulate the whole thing and see what happens, but there’s no reason to expect any deeper understanding. Organisms did not evolve to be understandable by humans.

I used to agree with this position. I used to argue that there was no reason to expect human-intelligible structure inside biological organisms, or deep neural networks, or other systems not designed to be understandable. But over the next few years after that biologist’s talk, I changed my mind, and one major reason for the change is Uri Alon’s book An Introduction to Systems Biology: Design Principles of Biological Circuits.

Alon's book is the ideal counterargument to the idea that organisms are inherently human-opaque: it directly demonstrates the human-understandable structures which comprise real biological systems. Right from the first page of the introduction:

> … one can, in fact, formulate general laws that apply to biological networks. Because it has evolved to perform functions, biological circuitry is far from random or haphazard. ... Although evolution works by random tinkering, it converges again and again onto a defined set of circuit elements that obey general design principles.
>
> The goal of this book is to highlight some of the design principles of biological systems... The main message is that biological systems contain an inherent simplicity. Although cells evolved to function and did not evolve to be comprehensible, simplifying principles make biological design understandable to us.

It's hard to update one's gut-level instinct that biology is a giant mess of spaghetti without seeing the structure first hand, so the goal of this post is to present just enough of the book to provide some intuition that, just maybe, biology really is human-understandable.

This review is prompted by the release of the book's second edition, just this past August, and that's the edition I'll follow through. I will focus specifically on the parts I find most relevant to the central message: biological systems are not opaque. I will omit the last three chapters entirely, since they have less of a gears-level focus and more of an evolutionary focus, although I will likely make an entire separate post on the last chapter (evolution of modularity).

Chapters 1-4: Bacterial Transcription Networks and Motifs

E-coli has about 4500 proteins, but most of those are chunked together into chemical pathways which work together to perform specific functions. Different pathways need to be expressed depending on the environment - for instance, e-coli won't express their lactose-metabolizing machinery unless the environment contains lots of lactose and not much glucose (which they like better).

In order to activate/deactivate certain genes depending on environmental conditions, bacteria use transcription factors: proteins sensitive to specific conditions, which activate or repress transcription of genes. We can think of the transcription factor activity as the cell's internal model of its environment. For example, from Alon:

> Many different situations are summarized by a particular transcription factor activity that signifies "I am starving". Many other situations are summarized by a different transcription factor activity that signifies "My DNA is damaged". These transcription factors regulate their target genes to mobilize the appropriate protein responses in each case.

The entire state of the transcription factors - the e-coli's whole model of its environment - has about 300 degrees of freedom. That's 300 transcription factors, each capturing different information, and regulating about 4500 protein genes.

Transcription factors often regulate the transcription of other transcription factors. This allows information processing in the transcription factor network. For instance, if either of two different factors (X, Y) can block transcription of a third (Z), then that's effectively a logical NOR gate: Z levels will be high when neither X nor Y is high. In general, transcription factors can either repress or promote (though rarely both), and arbitrarily complicated logic is possible in principle - including feedback loops.

Now we arrive at our first major piece of evidence that organisms aren't opaque spaghetti piles: bacterial transcription network motifs.

Random mutations form random connections between transcription factors - mutations can make any given transcription factor regulate any other very easily. But actual transcription networks do not look like random graphs. Here's a visualization from the book:

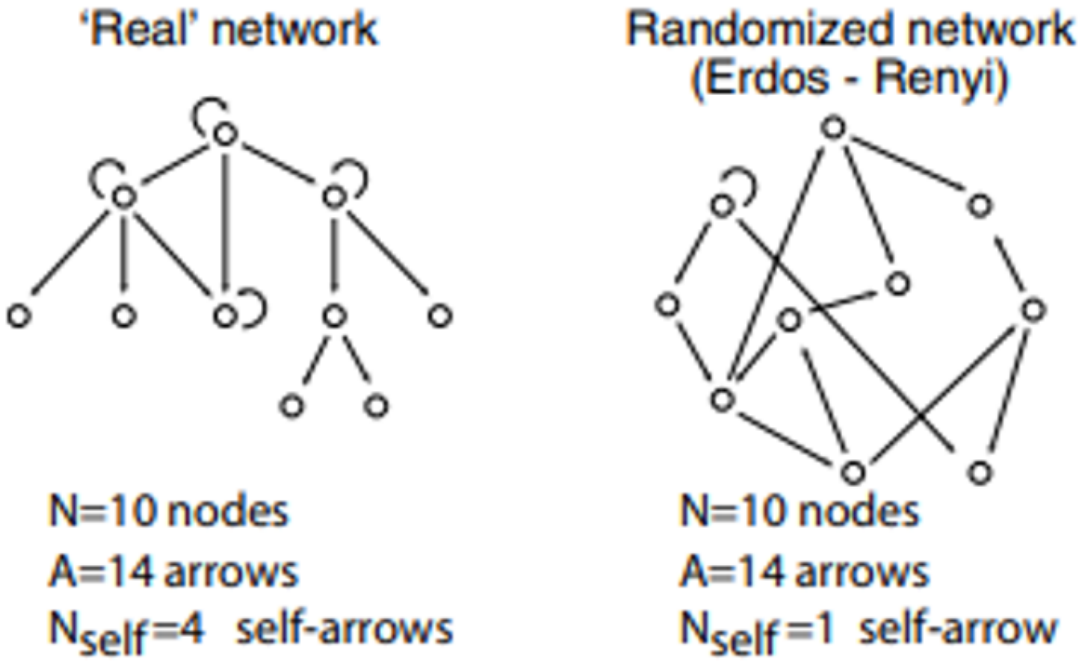


FIGURE 2.2

A few differences are immediately visible:

- Real networks have much more autoregulation (transcription factors activating/repressing their own transcription) than random networks
- Other than self-loops (aka autoregulation), real networks contain almost no feedback loops (at least in bacteria), though such loops are quite common in random networks
- Real networks are mostly tree-shaped; most nodes have at most a single parent.

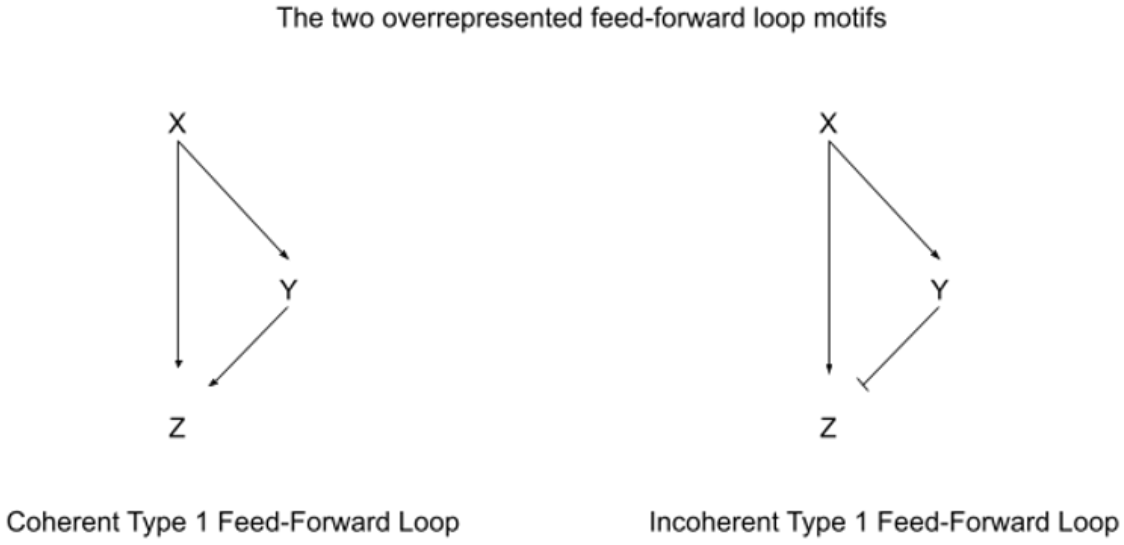
These patterns can be quantified and verified statistically via "motifs" (or "antimotifs"): connection patterns which occur much more frequently (or less frequently) in real transcription factor networks than in random networks.

Alon uses an e-coli transcription network with 424 nodes and 519 connections to quantify motifs. Chapters 2-4 each look at a particular class of motifs in detail:

- Chapter 2 looks at autoregulation. If the network were random, we'd expect about 1.2 ± 1.1 autoregulatory loops. The actual network has 40.
- Chapter 3 looks at three-node motifs. There is one massively overrepresented motif: the feed-forward loop (see diagram below), with 42 instances in the real network and only 1.7 ± 1.3 in a random network. Distinguishing activation from repression, there are eight possible feed-forward loop types, and two of the eight account for 80% of the feed-forward loops in the real network.

- Chapter 4 looks at larger motifs, though it omits the statistics. Fan-in and fan-out patterns, as well as fanned-out feed-forward loops, are analyzed.

Alon analyzes the chemical dynamics of each pattern, and discusses what each is useful for in a cell - for instance, autoregulatory loops can fine-tune response time, and feed-forward loops can act as filters or pulse generators.



Chapters 5-6: Feedback and Motifs in Other Biological Networks

Chapter 5 opens with developmental transcription networks, the transcription networks which lay out the body plan and differentiate between cell types in multicellular organisms. These are somewhat different from the bacterial transcription networks discussed in the earlier chapters. Most of the overrepresented motifs in bacteria are also overrepresented in developmental networks, but there are also new overrepresented motifs - in particular, positive autoregulation and two-node positive feedback.

Both of these positive feedback patterns are useful mainly for inducing bistability - i.e. multiple stable steady states. A bistable system with steady states A and B will stay in A if it starts in A, or stay in B if it starts in B, meaning that it can be used as a stable memory element. This is especially important to developmental systems, where cells need to decide what type of cell they will become (in coordination with other cells) and then stick to it - we wouldn’t want a proto-liver cell changing its mind and becoming a proto-kidney cell instead.

After discussing positive feedback, Alon includes a brief discussion of motifs in other biological networks, including protein-protein interactions and neuronal networks. Perhaps surprisingly (especially for neuronal networks), these include many of the same overrepresented motifs as transcription factor networks - suggesting universal principles at work.

Finally, chapter 6 is devoted entirely to biological oscillators, e.g. circadian rhythms or cell-cycle regulation or heart beats. The relevant motifs involve negative feedback loops. The main surprise is that oscillations can sometimes be sustained even when it seems like they should die out over time - thermodynamic noise in chemical concentrations can “kick” the system so that the oscillations continue indefinitely.

At this point, the discussion of motifs in biological networks wraps up. Needless to say, plenty of references are given which quantify motifs in various biological organisms and network types.

Chapters 7-8: Robust Recognition and Signal-Passing

There's quite a bit of hidden purpose in biological systems - seemingly wasteful side-reactions or seemingly arbitrary reaction systems turn out to be functionally critical. Chapters 7-8 show that robustness is one such "hidden" purpose: biological systems are buffeted by thermodynamic noise, and their functions need to be robust to that noise. Once we know to look for it, robustness shows up all over, and many seemingly-arbitrary designs don't look so random anymore.

Chapter 7 mainly discusses kinetic proofreading, a system used by both ribosomes (RNA-reading machinery) and the immune system to reduce error rates. At first glance, kinetic proofreading just looks like a wasteful side-reaction: the ribosome/immune cell binds its target molecule, then performs an energy-consuming side reaction and just waits around a while before it can move on to the next step. And if the target unbinds at any time, then it has to start all over again!

Yet this is exactly what's needed to reduce error rates.

The key is that the *correct* target is always most energetically stable to bind, so it stays bound longer (on average) than *incorrect* targets. At equilibrium, maybe 1% of the bound targets are incorrect. The irreversible side-reaction acts as a timer: it marks that some target is bound, and starts time. If the target falls off, then the side-reaction is undone and the whole process starts over… but the incorrect targets fall off much more quickly that the correct targets. So, we end up with correct targets "enriched": the fraction of incorrect targets drops well below its original level of 1%. Both the delay and the energy consumption are necessary in order for this to work: the delay to give the incorrect targets time to fall off, and the energy consumption to make the timer irreversible (otherwise everything just equilibrates back to 1% error).

Alon offers an analogy, in which a museum curator wants to separate the true Picasso lovers from the non-lovers. The Picasso room usually has about 10x more lovers than non-lovers (since the lovers spend much more time in the room), but the curator wants to do better. So, with a normal mix of people in there, he locks the incoming door and opens a one-way door out. Over the next few minutes, only a few of the picasso lovers leave, but practically all the non-lovers leave - Picasso lovers end up with much more than the original 10x enrichment in the room. Again, we see both key pieces: irreversibility and a delay.

It's also possible to stack such systems, performing multiple irreversible side-reactions in sequence, in order to further lower the error rate. Alon goes into much more depth, and explains the actual reactions involved in more detail.

Chapter 8 then dives into a different kind of robustness: robust signal-passing. The goal here is to pass some signal from outside the cell to inside. The problem is, there's a lot of thermodynamic noise in the number of receptors - if there happen to be 20% more receptors than average, then a simple detection circuit would measure 20% stronger signal. This problem can be avoided, but it requires a specific - and nontrivial - system structure.

In this case, the main trick is to have the receptor both activate and deactivate (i.e. phosphorylate and dephosphorylate) the internal signal molecule, with rates depending on whether the receptor is bound. At first glance, this might seem wasteful: what's the point of a receptor which undoes its own effort? But for robustness, it's critical - because the receptor both activates and deactivates the internal signal, its concentration cancels out in the

equilibrium expression. That means that the number of receptors won't impact the equilibrium activity level of the signal molecule, only how fast it reaches equilibrium.

The trick can also be extended to provide robustness to the background level of the signal molecule itself - Alon provides more detail. As you might expect, this type of structure is a common pattern in biological signal-receptor circuits.

For our purposes, the main takeaway from these two chapters is that, just because the system *looks* wasteful/arbitrary, does not mean it *is* . Once we know what to look for, it becomes clear that the structure of biological systems is not nearly so arbitrary as it looks.

Chapters 9-11: Exact Adaptation, Fold Change and Related Topics

When we move from an indoor room into full sunlight, our eyes quickly adjust to the brightness. A bacteria swimming around in search of food can detect chemical gradients among background concentrations varying by three orders of magnitude. Beta cells in the pancreas regulate glucose usage, bringing the long-term blood glucose concentration back to 5 mM, even when we shift to eating or exercising more. In general, a wide variety of biological sensing systems need to be able to detect changes and then return to a stable baseline, across a wide range of background intensity levels.

Alon discusses three problems in this vein, each with its own chapter:

- Exact adaptation: the "output signal" of a system always returns to the same baseline when the input stops changing, even if the input settles at a new level.
- Fold change: the system responds to percentage changes, across several decibels of background intensity.
- Extracellular versions of the above problems, in which control is decentralized.

Main takeaway: fairly specific designs are needed to achieve robust behavior.

Exact Adaptation

The main tool used for exact adaptation will be immediately familiar to engineers who've seen some linear control theory: integral feedback control. There are three key pieces:

- Some internal state variable M - e.g. concentration/activation of some molecule type or count of some cell type - used to track "error" over time
- An "internal" signal X
- An "external" signal and a receptor, which increases production/activation of the internal signal whenever it senses the external signal

The "error" tracked by the internal state M is the difference between the internal signal's concentration X and its long-term steady-state concentration X^*. The internal state increases/decreases in direct proportion to that difference, so that over time, the M is proportional to the integral \int_t (X^* - X) dt. Then, M itself represses production/activation of the internal signal X.

The upshot: if the external signal increases, then at first the internal signal X also increases, as the external receptor increases production/activation of X. But this pushes X above its long-term steady-state X^*, so M gradually increases, repressing X. The longer and further X is above its steady-state, the more M increases, and the more X is repressed. Eventually, M reaches a level which balances the new average level of the external signal, and X returns to the baseline.

Alon then discusses robustness of this mechanism compared to other possible mechanisms. Turns out, this kind of feedback mechanism is robust to changes in the background level of M, X, etc - steady-state levels shift, but the qualitative behavior of exact adaptation remains. Other, “simpler” mechanisms do not exhibit such robustness.

Fold-Change Detection

Fold-change detection is a pretty common theme in biological sensory systems, from eyes to bacterial chemical receptors. Weber’s Law is the general statement: sensory systems usually respond to changes on a log scale.

There’s two important pieces here:

- “Respond to changes” means exact adaption - the system returns to a neutral steady-state value in the long run when nothing is changing.
- “Log scale” means it’s percent changes which matter, and the system can work across several orders of magnitude of external signal

Alon gives an interesting example: apparently if you use a screen and an eye-tracker to cancel out a person’s rapid eye movements, their whole field of vision turns to grey and they can’t see anything. That’s responding to changes. On the other hand, if we step into bright light, background intensity can easily jump by an order of magnitude - yet a 10% contrast looks the same in low light or bright light. That’s operating on a log-scale.

Again, there’s some pretty specific criteria for systems to exhibit fold-change detection - few systems have consistent, useful behavior over multiple orders of magnitude of input values. Alon gives two particular circuits, as well as a general criterion.

Extracellular/Decentralized Adaptation

Alon moves on to the example of blood glucose regulation. Blood glucose needs to be kept at a pretty steady 5 mM level long-term; too low will starve the brain, and too high will poison the brain. The body uses an integral feedback mechanism to achieve robust exact adaptation of glucose levels, with the count of pancreatic beta cells serving as the state variable: when glucose is too low, the cells (slowly) die off, and when glucose is too high, the cells (slowly) proliferate.

The main new player is insulin. Beta cells do not themselves produce or consume much glucose; rather, they produce insulin, which we can think of as an inverse-price signal for glucose. When insulin levels are low (so the “price” of glucose is high), many cells throughout the body cut back on their glucose consumption. The beta cells serve as market-makers: they adjust the insulin/price level until the glucose market clears - meaning that there is no long-term increase or decrease in blood glucose.

A very similar system exists for many other metabolite/hormone pairs. For instance, calcium and parathyroid uses a nearly-identical system: integral feedback mechanism using cell count as a state variable with a hormone serving as price-signal to provide decentralized feedback control throughout the body.

Alon also spends a fair bit of time on one particular issue with this set-up: mutant cells which mismeasure the glucose concentration could proliferate and take over the tissue. One defense against this problem is for the beta cells to die when they measure very high glucose levels (instead of proliferating very quickly). This handles must mutations, but it also means that sufficiently high glucose levels can trigger an unstable feedback loop: beta cells die, which reduces insulin, which means higher glucose “price” and less glucose usage throughout the body, which pushes glucose levels even higher. That’s type-2 diabetes.

Chapter 12: Morphological Patterning

The last chapter we'll cover here is on morphological patterning: the use of chemical reactions and diffusion to lay out the body plans of multicellular organisms.

The basic scenario involves one group of cells (A) producing some signal molecule, which diffuses into a neighboring group of cells (B). The neighbors then differentiate themselves based on how strong the signal is: those nearby A will see high signal, so they adopt one fate, while those farther away see lower signal, so they adopt another fate, with some cutoff in between.

This immediately runs into a problem: if A produces too much or too little of the signal molecule, then the cutoff will be too far to one side or the other - e.g. the organism could end up with a tiny rib and big space between ribs, or a big rib and a tiny space between. It's not robust.

Once again, the right design can mitigate the problem.

Apparently one group ran a brute-force search over parameter space, looking for biologically-plausible systems which produced robust patterning. Only a few tiny corners of the parameter space worked, and those tiny corners all used a qualitatively similar mechanism. Alon explains the mechanism in some depth, but I'm gonna punt on it - much as I enjoy nonlinear PDEs (and this one is even analytically tractable), I'm not going to inflict them on readers here.

Once again, though it may seem that evolution can solve problems a million different ways and it's hopeless to look for structure, it actually turns out that only a few specific designs work - and those are understandable by humans.

Takeaway

Let's return to the Alon quote from the introduction:

> Because it has evolved to perform functions, biological circuitry is far from random or haphazard. ... Although evolution works by random tinkering, it converges again and again onto a defined set of circuit elements that obey general design principles.
>
> The goal of this book is to highlight some of the design principles of biological systems... The main message is that biological systems contain an inherent simplicity. Although cells evolved to function and did not evolve to be comprehensible, simplifying principles make biological design understandable to us.

We've now seen both general evidence and specific examples of this.

In terms of general evidence, we've seen that biological regulatory networks do not look statistically random. Rather, a handful of patterns - "motifs" - repeat often, lending the system a lot of consistent structure. Even though the system was not designed to be understandable, there's still a lot of recognizable internal structure.

In terms of specific examples, we've seen that only a small subset of possible designs can achieve certain biological goals:

- Robust recognition of molecules
- Robust signal-passing
- Robust exact adaptation and distributed exact adaptation
- Fold-change detection

- Robust morphological patterning

The designs which achieve robustness are exactly the designs used by real biological systems. Even though the system was not designed to be understandable, the simple fact that it works *robustly* forces the use of a handful of understandable structures.

A final word: when we do not understand something, it does not *look* like there is anything to be understood at all - it just looks like random noise. Just because it looks like noise does not mean there is no hidden structure.

Chris Olah’s views on AGI safety

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

Note: I am not Chris Olah. This post was the result of lots of back-and-forth with Chris, but everything here is my interpretation of what Chris believes, not necessarily what he actually believes. Chris also wanted me to emphasize that his thinking is informed by all of his colleagues on the OpenAI Clarity team and at other organizations.

In thinking about AGI safety—and really any complex topic on which many smart people disagree—I’ve often found it very useful to build a collection of different viewpoints from people that I respect that I feel like I understand well enough to be able to think from their perspective. For example, I will often try to compare what an idea feels like when I put on my Paul Christiano hat to what it feels like when I put on my Scott Garrabrant hat. Recently, I feel like I’ve gained a new hat that I’ve found extremely valuable that I also don’t think many other people in this community have, which is my Chris Olah hat. The goal of this post is to try to give that hat to more people.

If you’re not familiar with him, Chris Olah leads the Clarity team at OpenAI and formerly used to work at Google Brain. Chris has been a part of many of the most exciting ML interpretability results in the last five years, including Activation Atlases, Building Blocks of Interpretability, Feature Visualization, and DeepDream. Chris was also a coauthor of “Concrete Problems in AI Safety.”

He also thinks a lot about technical AGI safety and has a lot of thoughts on how ML interpretability work can play into that—thoughts which, unfortunately, haven’t really been recorded previously. So: here’s my take on Chris’s AGI safety worldview.

The benefits of transparency and interpretability

Since Chris primarily works on ML transparency and interpretability, the obvious first question to ask is how he imagines that sort of research aiding with AGI safety. When I was talking with him, Chris listed four distinct ways in which he thought transparency and interpretability could help, which I’ll go over in his order of importance.

Catching problems with auditing

First, Chris says, interpretability gives you a mulligan. Before you deploy your AI, you can throw all of your interpretability tools at it to check and see what it actually learned and make sure it learned the right thing. If it didn’t—if you find that it’s learned some sort of potentially dangerous proxy, for example—then you can throw your AI out and try again. As long as you’re in a domain where your AI isn’t actively trying to deceive your interpretability tools (via deceptive alignment, perhaps), this sort of a mulligan could help quite a lot in resolving more standard robustness problems (proxy alignment, for example). That being said, that doesn’t necessarily mean waiting until you’re on the verge of deployment to look for flaws. Ideally you’d be able to discover

problems early on via an ongoing auditing process as you build more and more capable systems.

One of the OpenAI Clarity team's major research thrusts right now is developing the ability to more rigorously and systematically audit neural networks. The idea is that interpretability techniques shouldn't have to "get lucky" to stumble across a problem, but should instead reliably catch any problematic behavior. In particular, one way in which they've been evaluating progress on this is the "auditing game." In the auditing game, one researcher takes a neural network and makes some modification to it—maybe images containing both dogs and cats are now classified as rifles, for example—and another researcher, given only the modified network, has to diagnose the problem and figure out exactly what modification was made to the network using only interpretability tools without looking at error cases. Chris's hope is that if we can reliably catch problems in an adversarial context like the auditing game, it'll translate into more reliably being able to catch alignment issues in the future.

Deliberate design

Second, Chris argues, advances in transparency and interpretability could allow us to significantly change the way we design ML systems. Instead of a sort of trial-and-error process where we just throw lots of different techniques at the various benchmarks and see what sticks, if we had significantly better transparency tools we might be able to design our systems deliberately by understanding why our models work and how to improve them. In this world, because we would be building systems with an understanding of why they work, we might be able to get a much better understanding of their failure cases as well and how to avoid them.

In addition to these direct benefits, Chris expects some large but harder-to-see benefits from such a shift as well. Right now, not knowing anything about how your model works internally is completely normal. If even partly understanding one's model became normal, however, then the amount we don't know might become glaring and concerning. Chris provides the following analogy to illustrate this: if the only way you've seen a bridge be built before is through unprincipled piling of wood, you might not realize what there is to worry about in building bigger bridges. On the other hand, once you've seen an example of carefully analyzing the structural properties of bridges, the absence of such an analysis would stand out.

Giving feedback on process

Third, access to good transparency and interpretability tools lets you give feedback to a model—in the form of a loss penalty, reward function, etc.—not just on its output, but also on the process it used to get to that output. Chris and his coauthors lay this argument out in "Building Blocks of Interpretability:"

> One very promising approach to training models for these subtle objectives is learning from human feedback. However, even with human feedback, it may still be hard to train models to behave the way we want if the problematic aspect of the model doesn't surface strongly in the training regime where humans are giving feedback. Human feedback on the model's decision-making process, facilitated by interpretability interfaces, could be a powerful solution to these problems. It might allow us to train models not just to make the right decisions, but to make them for the right reasons. (There is however a danger here: we are optimizing our model to

look the way we want in our interface — if we aren't careful, this may lead to the model fooling us!)

The basic idea here is that rather than just using interpretability as a mulligan at the end, you could also use it as part of your objective during training, incentivizing the model to be as transparent as possible. Chris notes that this sort of thing is quite similar to the way in which we actually judge human students by asking them to show their work. Of course, this has risks—it could increase the probability that your model only looks transparent but isn't actually—but it also has the huge benefit of helping your training process steer clear of bad uninterpretable models. In particular, I see this as potentially being a big boon for informed oversight, as it allows you to incorporate into your objective an incentive to be more transparent to an amplified overseer.

One way in particular that the Clarity team's work could be relevant here is a research direction they're working on called model diffing. The idea of model diffing is to have a way of systematically comparing different models and determining what's different from the point of view of high-level concepts and abstractions. In the context of informed oversight—or specifically relaxed adversarial training—you could use model diffing to track exactly how your model is evolving over the course of training in a way which is inspectable to the overseer.[1]

Building microscopes not agents

One point that Chris likes to talk about is that—despite talking a lot about how we want to avoid race-to-the-bottom dynamics—the AI safety community seems to have just accepted that we have to build agents, despite the dangers of agentic AIs.[2] Of course, there's a reason for this: agents seem to be more competitive. Chris cites Gwern's "Why Tool AIs Want to Be Agent AIs" here, and notes that he mostly agrees with it—it does seem like agents will be more competitive, at least by default.

But that still doesn't mean we have to build agents—there's no universal law compelling us to do so. Rather, agents only seem to be on the default path because a lot of the people who currently think about AGI see them as the shortest path.[3] But potentially, if transparency tools could be made significantly better, or if a major realignment of the ML community could be achieved—which Chris thinks might be possible, as I'll talk about later—then there might be another path.

Specifically, rather than using machine learning to build agents which directly take actions in the world, we could use ML as a *microscope*—a way of learning about the world without directly taking actions in it. That is, rather than training an RL agent, you could train a predictive model on a bunch of data and use interpretability tools to inspect it and figure out what it learned, then use those insights to inform—either with a human in the loop or in some automated way—whatever actions you actually want to take in the world.

Chris calls this alternative vision of what an advanced AI system might look like a microscope AI since the AI is being used sort of like a microscope to learn about and build models of the world. In contrast with something like a tool or oracle AI that is designed to output useful information, the utility of a microscope AI wouldn't come from its output but rather our ability to look inside of it and access all of the implicit knowledge it learned. Chris likes to explain this distinction by contrasting Google Translate—the oracle/tool AI in this analogy—to an interface that could give you access

to all the linguistic knowledge implicitly present in Google Translate—the microscope AI.

Chris talks about this vision in his post "Visualizing Representations: Deep Learning and Human Beings:"

> The visualizations are a bit like looking through a telescope. Just like a telescope transforms the sky into something we can see, the neural network transforms the data into a more accessible form. One learns about the telescope by observing how it magnifies the night sky, but the really remarkable thing is what one learns about the stars. Similarly, visualizing representations teaches us about neural networks, but it teaches us just as much, perhaps more, about the data itself.
>
> (If the telescope is doing a good job, it fades from the consciousness of the person looking through it. But if there's a scratch on one of the telescope's lenses, the scratch is highly visible. If one has an example of a better telescope, the flaws in the worse one will suddenly stand out. Similarly, most of what we learn about neural networks from representations is in unexpected behavior, or by comparing representations.)
>
> Understanding data and understanding models that work on that data are intimately linked. In fact, I think that understanding your model has to imply understanding the data it works on.
>
> While the idea that we should try to visualize neural networks has existed in our community for a while, this converse idea—that we can use neural networks for visualization—seems equally important [and] is almost entirely unexplored.

Shan Carter and Michael Nielsen have also discussed similar ideas in their Artificial Intelligence Augmentation article in Distill.

Of course, the obvious question with all of this is whether it could ever be anything but hopelessly uncompetitive. It is important to note that Chris generally agrees that microscopes are unlikely to be competitive—which is why he's mostly betting on the other routes to impact above. He just hasn't entirely given up hope that a realignment of the ML community away from agents towards things like deliberate design and microscopes might still be possible.

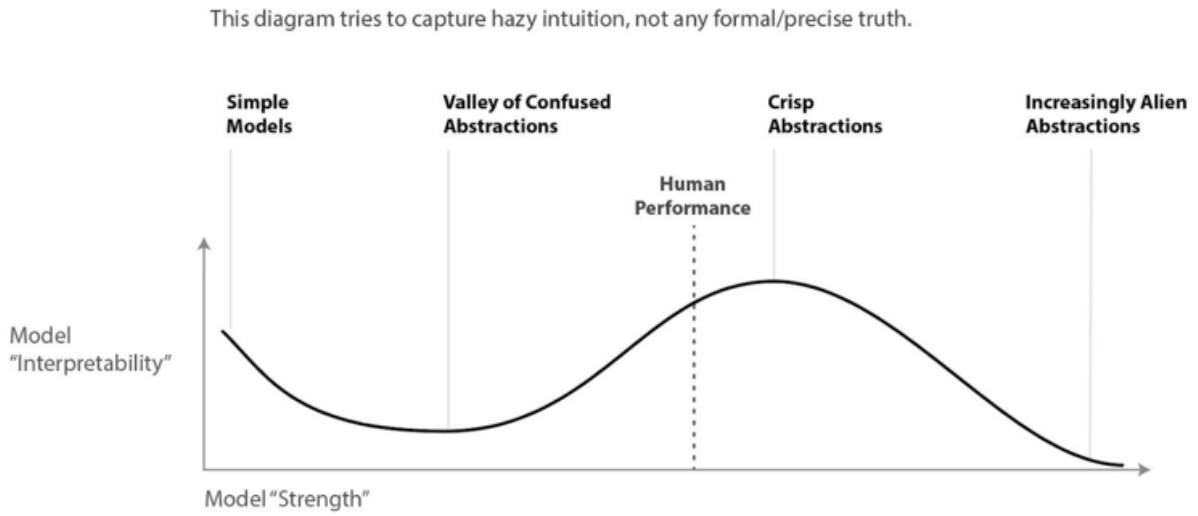
Furthermore, even in a world where the ML community still looks very similar to how it does today, if we have really good interpretability tools and the largest AI coalition has a strong lead over the next largest, then it might be possible to stick with microscopes for quite some time. Perhaps enough to either figure out how to align agents or otherwise get some sort of decisive strategic advantage.

What if interpretability breaks down as AI gets more powerful?

Chris notes that one of the biggest differences between him and many of the other people in the AI safety community is his belief that very strong interpretability is at all possible. The model that Chris has here is something like a reverse compilation process that turns a neural network into human-understable code. Chris notes that the resulting code might be truly gigantic—e.g. the entire Linux kernel—but that it would be faithful to the model and understandable by humans. Chris's basic intuition here is that neural

networks really do seem to learn meaningful features and that if you're willing to put a lot of energy in to understand them all—e.g. just actually inspect every single neuron—then you can make it happen. Chris notes that this is in contrast to a lot of other neural network interpretability work which is more aimed at approximating what neural networks do in particular cases.

Of course, this is still heavily dependent on exactly what the scaling laws are like for how hard interpretability will be as our models get stronger and more sophisticated. Chris likes to use the following graph to describe how he sees transparency and interpretability tools scaling up:



This graph has a couple of different components to it. First, simple models tend to be pretty interpretable—think for example linear regression, which gives you super easy-to-understand coefficients. Second, as you scale up past simple stuff like linear regression, things get a lot messier. But Chris has a theory here: the reason these models aren't very interpretable is because they don't have the capacity to express the full concepts that they need, so they rely on confused concepts that don't quite track the real thing. In particular, Chris notes that he has found that better, more advanced, more powerful models tend to have crisper, clearer, more interpretable concepts—e.g. InceptionV1 is more interpretable than AlexNet. Chris believes that this sort of scaling up of interpretability will continue for a while until you get to around human-level performance, at which point Chris hypothesizes that the trend will stop as models start moving away from crisp human-level concepts to still crisp but now quite alien concepts.

If you buy this graph—or something like it—then interpretability should be pretty useful all the way up to and including AGI—though perhaps not for very far past AGI. But if you buy a continuous-takeoff worldview, then that's still pretty useful. Furthermore, in my opinion, I think that the dropping off of interpretability at the end of this graph is just an artifact of using a human overseer. If you instead substituted in an amplified overseer, then I think it's plausible that interpretability could just keep going up, or at least level off at some high level.

Improving the field of machine learning

One thing that Chris thinks could really make a big difference in achieving a lot of the above goals would be some sort of realignment of the machine learning community. Currently, the thing that the ML community primarily cares about is chasing state-of-the-art results on its various benchmarks without regard for understanding what the ML tools they're using are actually doing. But that's not what the machine learning discipline has to look like, and in fact, it's not what most scientific disciplines do look like.

Here's Chris's vision for what an alternative field of machine learning might look like. Currently, machine learning researchers primarily make progress on benchmarks via trial and error. Instead, Chris wants to see a field which focuses on deliberate design where understanding models is prioritized and the way that people make progress is through deeply understanding their systems. In this world, ML researchers primarily make better models by using interpretability tools to understand why their models are doing what they're doing instead of just throwing lots of things at the wall and seeing what sticks. Furthermore, a large portion of the field in this world is just devoted to gathering information on what models do—cataloging all the different types of circuits that appear across different neural networks, for example[4]—rather than on trying to build new models.[5]

If you want to change the field in this way, there are essentially two basic paths to making something like that happen—you can either:

1. get current ML researchers to switch over to interpretability/deliberate design/microscope use or
2. produce new ML researchers working on those things.

Chris has thoughts on how to do both of these, but I'll start with the first one. Chris thinks that several factors could make a high-quality interpretability field appealing for researchers. First, interpretability could be a way for researchers without access to large amounts of compute to stay relevant in a world where relatively few labs can train the largest machine learning models. Second, Chris thinks there's lots of low hanging fruit in interpretability such that it should be fairly easy to have impressive research results in the space over the next few years. Third, Chris's vision of interpretability is very aligned with traditional scientific virtues—which can be quite motivating for many people—even if it isn't very aligned with the present paradigm of machine learning.

However, If you want researchers to switch to a new research agenda and/or style of research, it needs to be possible for them to support careers based on it. Unfortunately, the unit of academic credit in machine learning tends to be traditional papers, published in conferences, evaluated on whether they set a new state-of-the-art on a benchmark (or more rarely by proving theoretical results). This is what decides who gets hired, promoted, and tenured in machine learning.

To address this, Chris founded Distill, an academic machine learning journal that aims to promote a different style of machine learning research. Distill aims to be a sort of "adapter" between the traditional method of evaluating research and the new style of research—based around things like deliberate design and microscope use—that Chris

wants to see the field move to. Specifically, Distill does this by being different in a few key ways:

1. Distill explicitly publishes papers visualizing machine learning systems, or even just explanations improving Clarity of thought in machine learning (Distill's expository articles have become widely used references).
2. Distill has all of the necessary trappings to make it recognized as a legitimate academic journal such that Distill publications will be taken seriously and cited.
3. Distill has support for all the sorts of nice interactive diagrams that are often necessary for presenting interpretability research.

The second option is to produce new ML researchers pursuing deliberate design rather than converting old ones. Here, Chris has a pretty interesting take on how this can be done: convert neuroscientists and systems biologists.

Here's Chris's pitch. There are whole fields of neuroscience dedicated to understanding all the different connections, circuits, pathways, etc. in all different manner of animal brains. Similarly, for the systems biologists, there are significant communities of researchers studying individual proteins, their interactions and pathways, etc. While neural networks are different from these lines of research at a detailed level, a lot of high level research expertise—e.g. epistemic standards for studying circuits, recurring motifs, research intuition—may be just as helpful for this type of research as machine learning expertise. Chris thinks neuroscientists or systems biologists willing to make this transition would be able to get funding to do their research, a much easier time running experiments, and lots of low-hanging fruit in terms of new publishable results that nobody has found yet.

Doesn't this speed up capabilities?

Yes, it probably does—and Chris agrees that there's a negative component to that—but he's willing to bet that the positives outweigh the negatives.

Specifically, Chris thinks the main question is whether principled and deliberate model design based on interpretability can beat automated model design approaches like neural architecture search. If it can, we get capabilities acceleration, but also a paradigm shift towards deliberate model design, which Chris expects to significantly aid alignment. If we don't, interpretability loses one of its upsides (other advantages like auditing still exist in this world) but also doesn't have the downside of acceleration. Both the upside and downside go hand in hand, and Chris expects the upside to outweigh the downside.

Update: If you're interested in understanding Chris's current transparency and interpretability work, a good starting point is the Circuits Thread on Distill.

---

1. In particular, this could be a way of getting traction on addressing gradient hacking. ↩

2. As an example of the potential dangers of agents, more agentic AI setups seem much more prone to mesa-optimization. ↩

3. A notable exception to this, however, is Eric Drexler's "Reframing Superintelligence: Comprehensive AI Services as General Intelligence." ↩

4. An example of the sort of common circuit that appears in lots of different models that the Clarity team has found is the way in which convolutional neural networks stay reflection-invariant: to detect a dog, they separately detect leftwards-facing and rightwards-facing dogs and then union them together. ↩

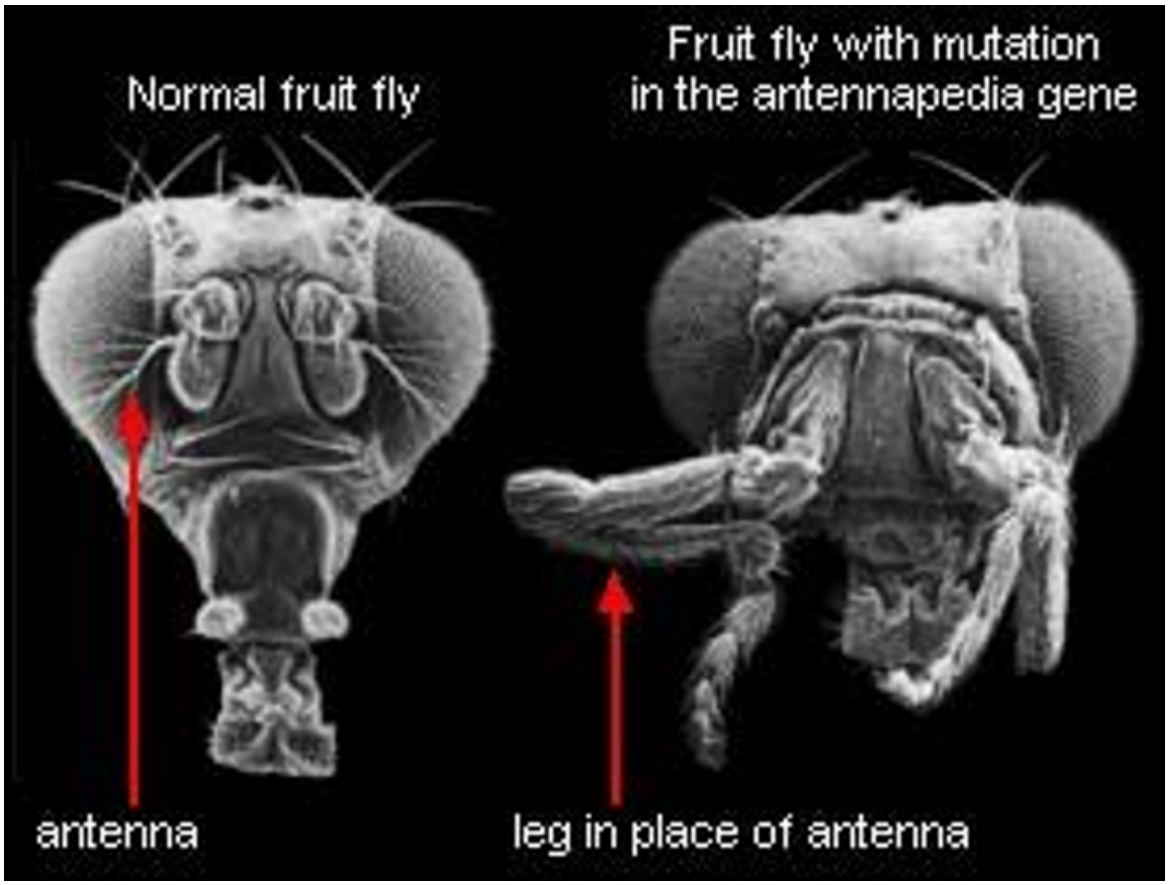
5. This results in a large portion of the field being focused on what is effectively microscope use, which could also be quite relevant for making microscope AIs more tenable. ↩

Evolution of Modularity

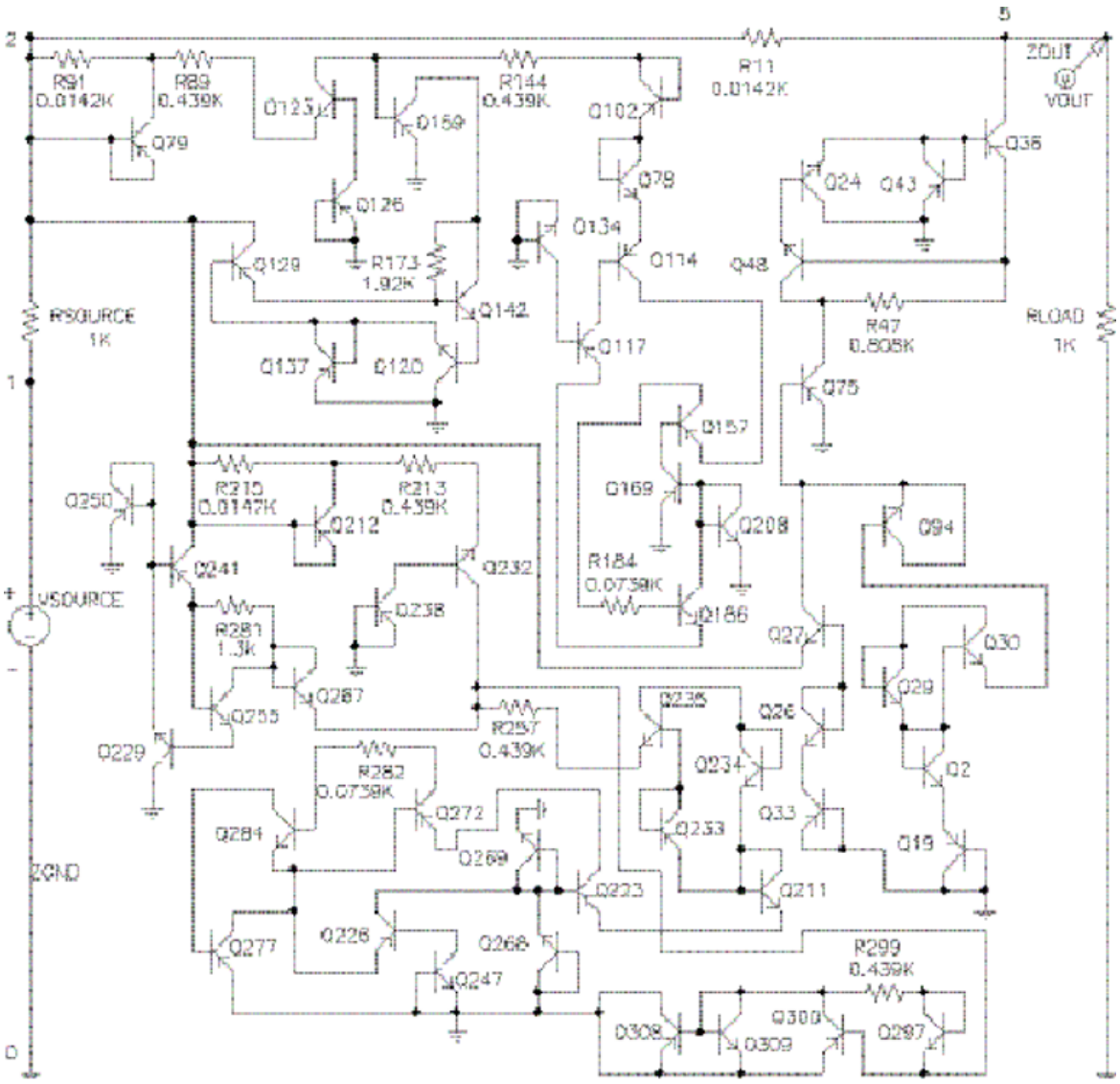
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is based on chapter 15 of Uri Alon's book An Introduction to Systems Biology: Design Principles of Biological Circuits. See the book for more details and citations; see here for a review of most of the rest of the book.

Fun fact: biological systems are highly modular, at multiple different scales. This can be quantified and verified statistically, e.g. by mapping out protein networks and algorithmically partitioning them into parts, then comparing the connectivity of the parts. It can also be seen more qualitatively in everyday biological work: proteins have subunits which retain their function when fused to other proteins, receptor circuits can be swapped out to make bacteria follow different chemical gradients, manipulating specific genes can turn a fly's antennae into legs, organs perform specific functions, etc, etc.



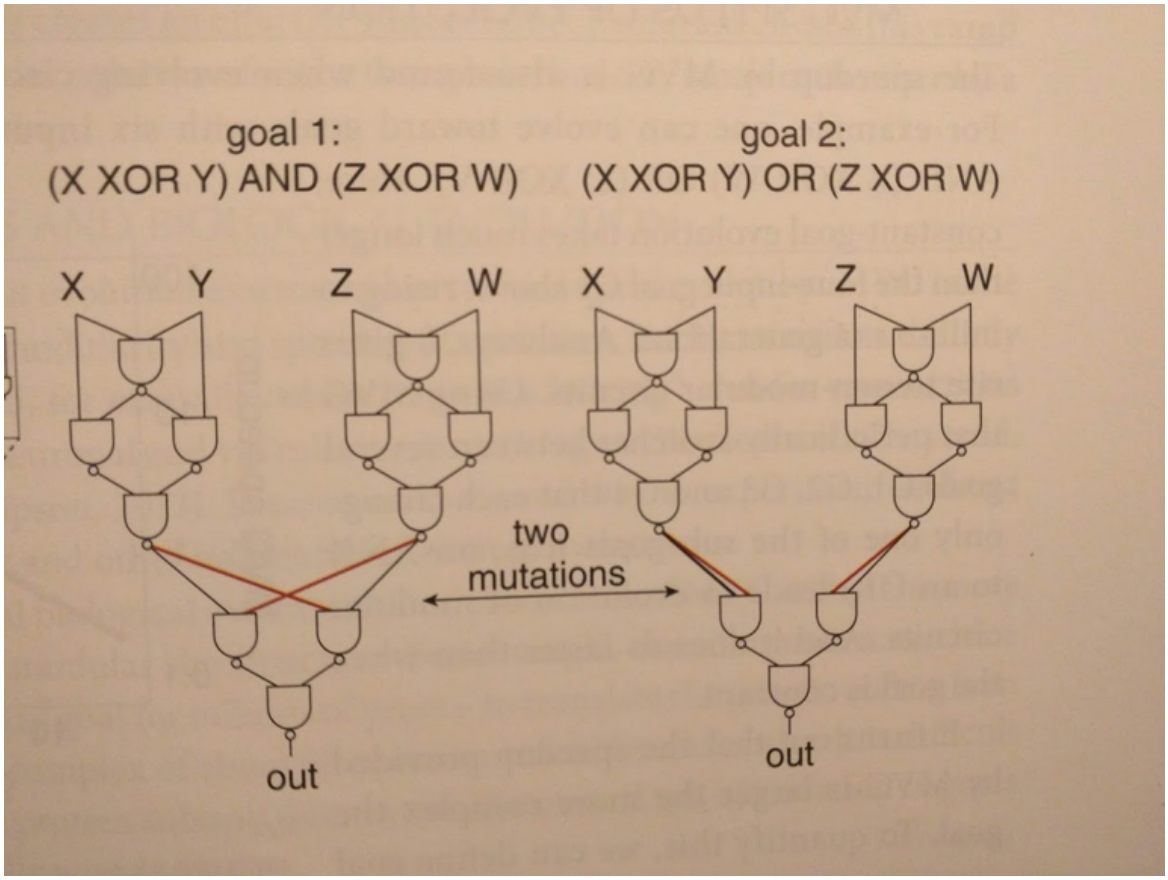
On the other hand, systems designed by genetic algorithms (aka simulated evolution) are decidedly *not* modular. This can also be quantified and verified statistically. Qualitatively, examining the outputs of genetic algorithms confirms the statistics: they're a mess.



So: what is the difference between real-world biological evolution vs typical genetic algorithms, which leads one to produce modular designs and the other to produce non-modular designs?

Kashtan & Alon tackle the problem by evolving logic circuits under various conditions. They confirm that simply optimizing the circuit to compute a particular function, with random inputs used for selection, results in highly non-modular circuits. However, they are able to obtain modular circuits using “modularly varying goals” (MVG).

The idea is to change the reward function every so often (the authors switch it out every 20 generations). Of course, if we just use completely random reward functions, then evolution doesn’t learn anything. Instead, we use “modularly varying” goal functions: we only swap one or two little pieces in the (modular) objective function. An example from the book:



The upshot is that our different goal functions generally use similar sub-functions - suggesting that they share sub-goals for evolution to learn. Sure enough, circuits evolved using MVG have modular structure, reflecting the modular structure of the goals.

(Interestingly, MVG also dramatically accelerates evolution - circuits reach a given performance level much faster under MVG than under a fixed goal, despite needing to change behavior every 20 generations. See either the book or the paper for more on that.)

How realistic is MVG as a model for biological evolution? I haven't seen quantitative evidence, but qualitative evidence is easy to spot. MVG as a theory of biological modularity predicts that highly variable subgoals will result in modular structure, whereas static subgoals will result in a non-modular mess. Alon's book gives several examples:

- Chemotaxis: different bacteria need to pursue/avoid different chemicals, with different computational needs and different speed/energy trade-offs, in various combinations. The result is modularity: separate components for sensing, processing and motion.
- Animals need to breathe, eat, move, and reproduce. A new environment might have different food or require different motions, independent of respiration or reproduction - or vice versa. Since these requirements vary more-or-less independently in the environment, animals evolve modular systems to deal with them: digestive tract, lungs, etc.
- Ribosomes, as an anti-example: the functional requirements of a ribosome hardly vary at all, so they end up non-modular. They have pieces, but most pieces do not have an obvious distinct function.

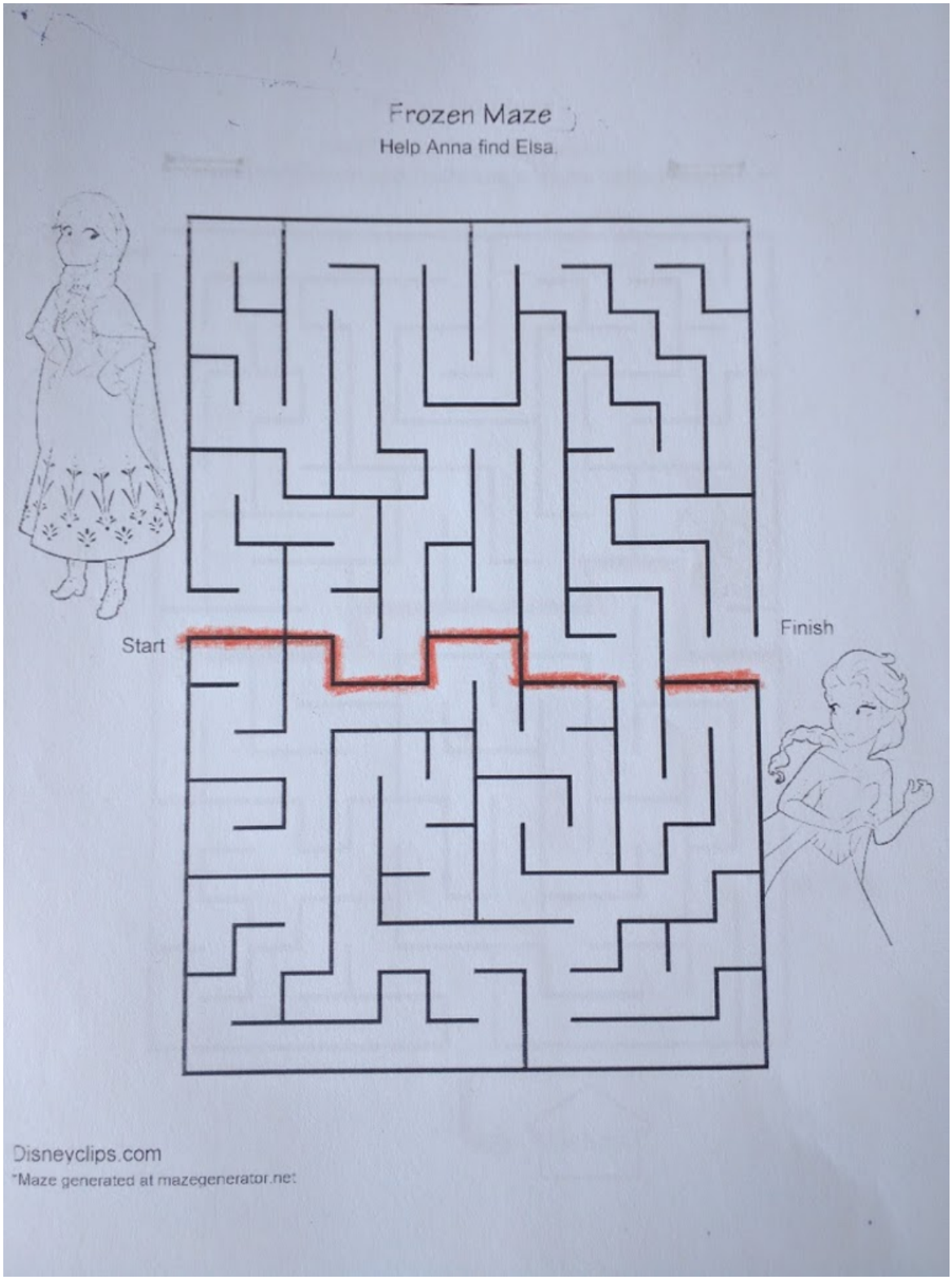
To sum it up: modularity in the system evolves to match modularity in the environment.

Gears-Level Models are Capital Investments

Mazes

The usual method to solve a maze is some variant of babble-and-prune: try a path, if it seems to get closer to the exit then keep going, if it hits a dead end then go back and try another path. It's a black-box method that works reasonably well on most mazes.

However, there are other methods. For instance, you could start by looking for a chain of walls with only one opening, like this:



This chain of walls is a gears-level insight into the maze - a piece of the internal structure which lets us better understand “how the maze works” on a low level. It’s not specific to any particular path, or to any particular start/end points - it’s a property of the maze itself. Every

shortest path between two points in the maze either starts and ends on the same side of that line, or passes through the gap.

If we only need to solve the maze once, then looking for a chain of walls is not very useful - it could easily take as long as solving the maze! But if we need to solve the *same* maze more than once, with different start and end points... then we can spend the time finding that chain of walls just once, and re-use our knowledge over and over again. It's a capital investment: we do some extra work up-front, and it pays out in lower costs every time we look for a path through the maze in the future.

This is a general feature of gears-level models: figuring out a system's gears takes extra work up-front, but yields dividends forever. The alternative, typically, is a black-box strategy: use a method which works without needing to understand the internals of the system. The black-box approach is cheaper for one-off tasks, but usually doesn't yield any insights which will generalize to new tasks using the same system - it's context-dependent.

Marketing

Suppose we work with the marketing team at an online car loan refinance company, and we're tasked with optimizing the company's marketing to maximize the number of car loans the company refinances. Here's two different approaches we might take:

- We a/b test hundreds of different ad spend strategies, marketing copy permutations, banner images, landing page layouts, etc. Ideally, we find a particular combination works especially well.
- We obtain some anonymized data from a credit agency on people with car loans. Ideally, we learn something about the market - e.g. maybe subprime borrowers usually either declare bankruptcy or dramatically increase their credit score within two years of taking a loan.

The first strategy is black-box: we don't need to know anything about who our potential customers are, what they want, the psychology of clicking on ads, etc. We can treat our marketing pipeline as a black box and fiddle with its inputs to see what works. The second strategy is gears-level, the exact opposite of black-box: the whole point is to learn who our potential customers are, breaking open the black box and looking at the internal gears.

These aren't mutually exclusive, and they have different relative advantages. Some upsides of black-box:

- Black-box is usually cheaper and easier, since the code involved is pretty standard and we don't need to track down external data. Gears-level strategies require more custom work and finding particular data.
- Black-box yields direct benefits when it works, whereas gears-level requires an extra step to translate whatever insights we find into actual improvements.

On the other hand:

- Gears-level insights can highlight ideas we wouldn't even have thought to try, whereas black-box just tests the things we think to test.
- When some tests are expensive (e.g. integrating with a new ad channel), gears-level knowledge can tell us which tests are most likely to be worthwhile.
- Black-box optimization is subject to Goodhart, while gears-level insights usually are not (at least in-and-of themselves)
- Gears-level insights are less likely subject to distribution shift. For instance, if we change ad channels, then the distribution of people seeing our ads will shift. Different ad copy will perform well, and we'd need to restart our black-box a/b testing, whereas general insights about subprime borrowers are more likely to remain valid.

- Conversely, black-box optimizations depreciate over time. Audiences and ad channels evolve, and ads need to change with them, requiring constant re-optimization to check that old choices are still optimal.
- By extension, gears-level insights tend to be permanent and broadly applicable, and have the potential for compound returns, whereas black-box improvements are much more context-specific and likely to shift with time.

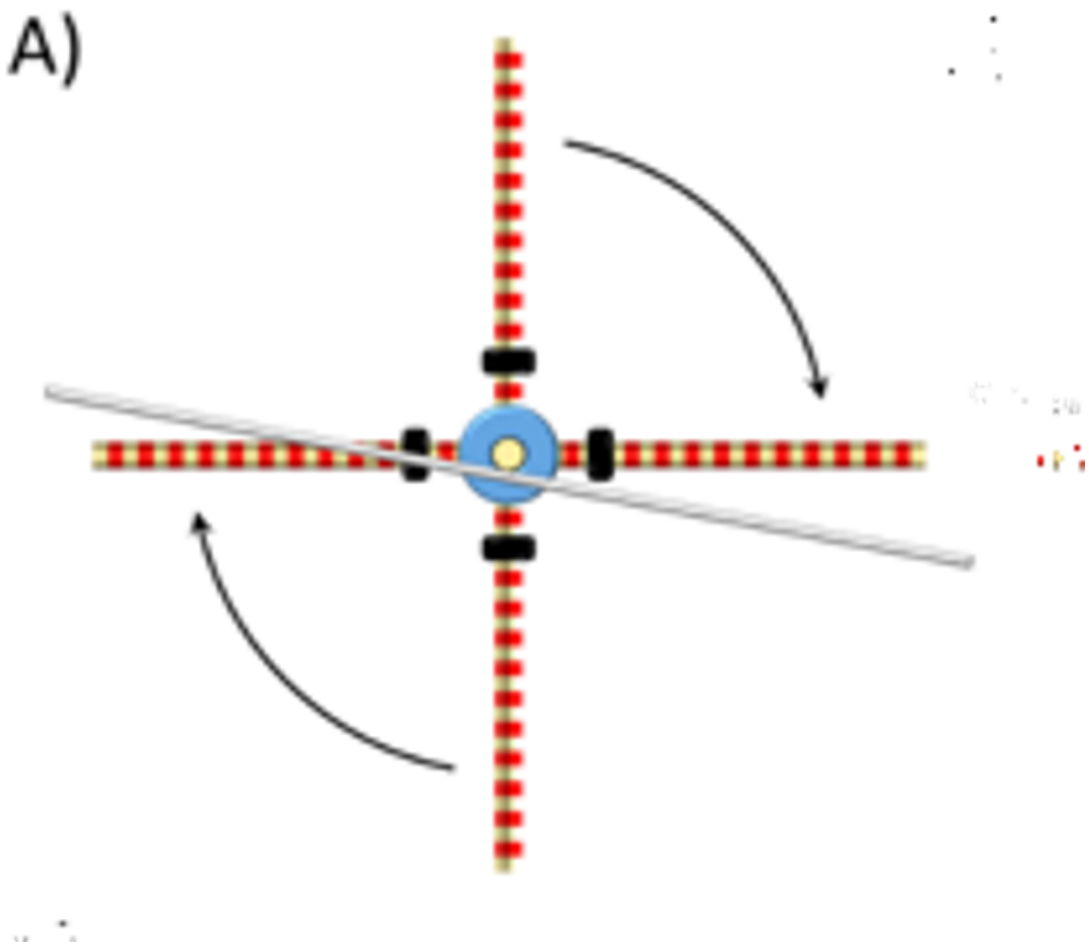
In short, the black-box approach is easier, cheaper, and more directly useful - but its benefits are ephemeral and it can't find unknown unknowns. Gears-level understanding is more difficult, expensive, and risky, but it offers permanent, generalizable insights and can suggest new questions we wouldn't have thought to ask.

With this in mind, consider the world through the eyes of an ancient lich or thousand-year-old vampire. It's a worldview in which ephemeral gains are irrelevant. All that matters is permanent, generalizable knowledge - everything else will fade in time, and usually not even very much time. In this worldview, gears-level understanding is everything.

On the other end of the spectrum, consider the world through the eyes of a startup with six months of runway which needs to show rapid growth in order to close another round of funding. For them, black-box optimization is everything - they want fast, cheap results which don’t need to last forever.

Wheel with Weights

There’s a neat experiment where people are given a wheel with some weights on it, each of which can be shifted closer to/further from the center. Groups of subjects have to cooperatively find settings for the weights which minimize the time for the wheel to roll down a ramp.



Given the opportunity to test things out, subjects would often iterate their way to optimal settings - but they didn't iterate their way to correct theories. When asked to predict how hypothetical settings would perform, subjects' predictions didn't improve much as they iterated. This is black-box optimization: optimization was achieved, but insight into the system was not.

If the problem had changed significantly - e.g. changing weight ratios/angles, ramp length/angle, etc - the optimal settings could easily change enough that subjects would need to re-optimize from scratch. On the other hand, the system is simple enough that just doing all the math is tractable - and that math would remain essentially the same if weights, angles, and lengths changed. A gears-level understanding is possible, and would reduce the cost of optimizing for new system parameters. It's a capital investment: it only makes sense to make the investment in gears-level understanding if it will pay off on many different future problems.

In the experiment, subjects were under no pressure to achieve gears-level understanding - they only needed to optimize for one set of parameters. I'd predict that people would be more likely to gain understanding if they needed to find optimal weight-settings quickly for many different wheel/ramp parameters. (A close analogy is evolution of modularity: changing objectives incentivize learning general structure.)

Metis

Let’s bring in the manioc example:

> There's this plant, manioc, that grows easily in some places and has a lot of calories in it, so it was a staple for some indigenous South Americans since before the Europeans showed up. Traditional handling of the manioc involved some elaborate time-consuming steps that had no apparent purpose, so when the Portuguese introduced it to Africa, they didn't bother with those steps - just, grow it, cook it, eat it.
>
> The problem is that manioc's got cyanide in it, so if you eat too much too often over a lifetime, you get sick, in a way that's not easily traceable to the plant. Somehow, over probably hundreds of years, the people living in manioc's original range figured out a way to leach out the poison, without understanding the underlying chemistry - so if you asked them why they did it that way, they wouldn't necessarily have a good answer.

The techniques for processing manioc are a stock example of metis: traditional knowledge accumulated over generations, which doesn’t seem like it has any basis in reason or any reason to be useful. It’s black-box knowledge, where the black-box optimizer is cultural transmission and evolution. Manioc is a cautionary tale about the dangers of throwing away or ignoring black-box knowledge just because it doesn’t contain any gears.

In this case, building a gears-level model was *very* expensive - people had to get sick on a large scale in order to figure out that any knowledge was missing at all, and even after that it presumably took a while for scientists to come along and link the problem to cyanide content. On the other hand, now that we have that gears-level model in hand, we can quickly and easily test new cooking methods to see whether they eliminate the cyanide - our gears-level model provides generalizable insights. We can even check whether any particular dish of manioc is safe before eating it, or breed new manioc strains which contain less cyanide. Metic knowledge would have no way to do any of that - it doesn’t generalize.

More Examples

(Note: in each of these examples, there are many other ways to formulate a black-box/gears-level approach. I just provide one possible approach for each.)

Pharma

- Black box approach: run a high-throughput assay to test the effect thousands of chemicals against low-level markers of some disease.
- Gears-level approach: comb the literature for factors related to some disease. Run experiments holding various subsets of the factors constant while varying others, to figure out which factors mediate the effect of which others, and ultimately build up a causal graph of their interactions.

The black-box approach is a lot cheaper and faster, but it’s subject to Goodhart problems, won’t suggest compounds that nobody thought to test, and won’t provide any knowledge which generalizes to related diseases. If none of the chemicals tested are effective, then the black-box approach leaves no foundation to build on. The gears-level approach is much slower and more expensive, but eventually yields reliable, generalizable knowledge.

Financial Trading

- Black box approach: build a very thorough backtester, then try out every algorithm or indicator we can think of to see if any of them achieve statistically significant improvement over market performance.
- Gears-level approach: research the trading algorithms and indicators actually used by others, then simulate markets with traders using those algorithms/indicators. Compare results against real price behavior and whatever side data can be found in order to identify missing pieces.

The gears-level approach is far more work, and likely won't produce anything profitable until very late in development. On the other hand, the gears-level approach will likely generalize far better to new markets, new market conditions, etc.

Data Science

- Black box approach: train a neural network, random forest, support vector machine, or whatever generic black-box learning algorithm you like.
- Gears-level approach: build a probabilistic graphical model. Research the subject matter to hypothesize model structure, and statistically compare different model structures to see which match the data best. Look for side information to confirm that the structure is correct.

The black box approach is subject to Goodhart and often fails to generalize. The gears-level approach is far more work, requiring domain expertise and side data and probably lots of custom code (although the recent surge of probabilistic programming languages helps a lot in that department), but gears-level models ultimately give us human-understandable explanations of how the system actually works. Their internal parameters have physical meaning.

Takeaway

Building gears-level models is expensive - often prohibitively expensive. Black-box approaches are usually much cheaper and faster. But black-box approaches rarely generalize - they're subject to Goodhart, need to be rebuilt when conditions change, don't identify unknown unknowns, and are hard to build on top of. Gears-level models, on the other hand, offer permanent, generalizable knowledge which can be applied to many problems in the future, even if conditions shift.

The upfront cost of gears-level knowledge makes it an investment, and the payoff of that investment is the ability to re-use the model many times in the future.

A mechanistic model of meditation

Meditation has been claimed to have all kinds of transformative effects on the psyche, such as improving concentration ability, healing trauma, cleaning up delusions, allowing one to track their subconscious strategies, and making one's nervous system more efficient. However, an explanation for why and how exactly this would happen has typically been lacking. This makes people reasonably skeptical of such claims.

In this post, I want to offer an explanation for one kind of a mechanism: meditation increasing the degree of a person's introspective awareness, and thus leading to increasing psychological unity as internal conflicts are detected and resolved.

Note that this post does *not* discuss "enlightenment". That is a related but separate topic. It is possible to pursue meditation mainly for its ordinary psychological benefits while being uninterested in enlightenment, and vice versa.

What is introspective awareness?

In an earlier post on introspective awareness, I distinguished between being *aware of something*, and *being aware of having been aware of something*. My example involved that of a robot whose consciousness contains one mental object at a time, and which is aware of different things at different times:

> Robot's thought at time 1: It's raining outside
>
> Robot's thought at time 2: Battery low
>
> Robot's thought at time 3: Technological unemployment protestors are outside
>
> Robot's thought at time 4: Battery low
>
> Robot's thought at time 5: I'm now recharging my battery

At times 2-5, the robot has no awareness of the fact that it was thinking about rain at time 1. As soon as something else captures its attention, it has no idea of this earlier conscious content - *unless* a particular subsystem happens to record the fact, and can later re-present the content in an appropriately tagged form:

> Time 6: At time 1, there was the thought that [It's raining outside]

I said that at time 6, the robot had a *moment of introspective awareness*: a mental object containing a summary of its previous thoughts, which can then be separately examined and acted upon.

Humans are not robots. But I previously summarized the neuroscience book *Consciousness and the Brain*, and its global neuronal workspace (GNW) model of consciousness. According to this model, the contents of consciousness correspond to what is being represented in a particular network of neurons - the global workspace - that connects different parts of the brain. Different systems are constantly competing to get their contents into the global workspace, which can only hold one piece of content at a time. Thus, like robots, we too are only aware of one thing at a time, and tend to lose awareness of our earlier thoughts - unless something reminds us of them.

In what follows, I will suggest that like robots, humans also have a type of conscious content that we might call *introspective awareness*, which allows us to be more aware of our previous mental activity. (I am borrowing the term from the meditation book *The Mind Illuminated,* which distinguishes between *introspective attention*, *introspective awareness*, and *metacognitive introspective awareness*. I am eliding these differences for the sake of simplicity.)

I will also explore the idea that introspective awareness is a sensory channel in a similar sense as vision and sound are. The experience of sight or sound is produced by subsystems which send information to consciousness; likewise, introspective awareness is produced by a subsystem which captures information in the brain and then sends it (back) to consciousness.

We can train our other senses to become more accurate and detailed. Gilbert, Sigman & Crist (2001), reviewing the neuroscience on sensory training, list a number of ways in which discrimination can be increased in a variety of sensory modalities: among other things, "visual acuity, somatosensory spatial resolution, discrimination of hue, estimation of weight, and discrimination of acoustical pitch all show improvement with practice"; even the spatial resolution of the visual system can be deliberately increased by training.

If introspective awareness is a sensory channel, can it also be practiced to improve the number of details it will pick up on? One may feel that I am stretching the metaphor here. But in fact, *Consciousness and the Brain* suggests that *all* sensory training is in a sense training in introspection. The additional information that we get by training our senses has always been collected by our brain, but that information has remained isolated at lower levels of processing. To make it conscious, one needs to grow new neural circuits which extract the lower-level information and re-encode it in a format which can be sent to consciousness.

Thus, the brain *already* has the ability to take normally unavailable subconscious information and make it consciously available by practice. What is needed is a way to point that learning process at the kind of information that we would normally consider "introspective", rather than on an external information source.

From *Consciousness and the Brain:*

> ... a fourth way in which neural information can remain unconscious, according to workspace theory, is to be diluted into a complex pattern of firing. To take a concrete example, consider a visual grating that is so finely spaced, or that flickers so fast (50 hertz and above), that you cannot see it. Although you perceive only a uniform gray, experiments show that the grating is actually encoded inside your brain: distinct groups of visual neurons fire for different orientations of the grating. Why can’t this pattern of neuronal activity be brought to consciousness? Probably because it makes use of an extremely tangled spatiotemporal pattern of firing in the primary visual area, a neural cipher too complex to be explicitly recognized by global workspace neurons higher up in the cortex. Although we do not yet fully understand the neural code, we believe that, in order to become conscious, a piece of information first has to be re-encoded in an explicit form by a compact assembly of neurons. The anterior regions of the visual cortex must dedicate specific neurons to meaningful visual inputs, before their own activity can be amplified and cause a global workspace ignition that brings the information into awareness. If the information remains diluted in the firing of myriad unrelated neurons, then it cannot be made conscious.

> Any face that we see, any word that we hear, begins in this unconscious manner, as an absurdly contorted spatiotemporal train of spikes in millions of neurons, each sensing only a minuscule part of the overall scene. Each of these input patterns contains virtually infinite amounts of information about the speaker, message, emotion, room size . . . if only we could decode it—but we can't. We become aware of this latent information only once our higher-level brain areas categorize it into meaningful bins. Making the message explicit is an essential role of the hierarchical pyramid of sensory neurons that successively extract increasingly abstract features of our sensations. Sensory training makes us aware of faint sights or sounds because, at all levels, neurons reorient their properties to amplify these sensory messages. Prior to learning, a neuronal message was already present in our sensory areas, but only implicitly, in the form of a diluted firing pattern inaccessible to our awareness.

Richard's therapy session

We saw an example of introspective awareness in my post on the book *Unlocking the Emotional Brain*. In the transcript, a man named Richard has been suffering from severe self-doubt, and is asked to imagine how it would feel like if he made confident comments in a work meeting. The following conversation follows:

> Richard: Now I'm feeling really uncomfortable, but-it's in a different way.
>
> Therapist: OK, let yourself feel it - this different discomfort. [Pause.] See if any words come along with this uncomfortable feeling.
>
> Richard: [Pause.] Now they hate me.

The therapist is asking Richard to focus his attention on the feeling of discomfort, generating moments of introspective awareness *about* the discomfort. Notice that Richard becomes more thoughtful and less reactive to the anxiety as he does so. My guess of what is happening is something like this:

When Richard is feeling anxious, this means that a mental object encoding something like "the feeling of anxiety" is being represented in the workspace. This activates neural rules which trigger the kinds of responses that anxiety has evolved to produce. For example, a system may be triggered which attempts to plan how to escape the situation causing the anxiety. This system's intentions are then injected into the workspace, producing a state of mind where the feeling of anxiety alternates with thoughts of how to get away.

Introspective awareness is its own type of mental object, produced by a different subsystem which takes inputs from the global workspace, re-encodes them in a format which highlights particular aspects of that data, and outputs that back into the workspace. When a representation of an anxious state of mind is created, that representation does not by itself trigger the same rules as the original anxiety did.

As a result, as representations of the anxiety begin to alternate *together with* the anxiety, there are proportionately less moments of anxiety. This in turn triggers fewer of the subsystems attempting to escape the situation, making it easier to reflect on the anxiety without being bothered by it.

When Richard's therapist asks him to feel the anxiety and to see if any words come along with it, the subsystem for introspective awareness was primed to look for any content that could be re-presented in verbal form. As Richard's anxiety had been produced by an emotional schema including a prediction that being confident makes you hated, some of that information had passed through the workspace and been available for the awareness subsystem to capture. This brought up the verbalization of what the schema predicted would happen if Richard was confident - "now they hate me".

> Therapist: "Now they hate me." Good. Keep going: See if this really uncomfortable feeling can also tell you why they hate you now.

According to the GNW model, when a particular piece of content is maintained as the center of attention, it strengthens the activation of any structures associated with it. As Richard's therapist guides him to focus on the verbal content, more information related to it is broadcast into the workspace. The further prompt guides the awareness subsystem to look for patterns that feel like the *reason* for the hate.

> Richard: [Pause.] Hnh. Wow. It's because… now I'm… an arrogant asshole… like my father… a totally self-centered, totally insensitive know-it-all.

The therapist then takes a pattern which Richard has brought up and helps crystallize it further, and throws it back to Richard for verification.

> Therapist: Do you mean that having a feeling of confidence as you speak turns you into an arrogant asshole, like Dad?
>
> Richard: Yeah, exactly. Wow.

In this example, we saw that having more moments of introspective awareness was beneficial for Richard. As aspects of his moment-to-moment consciousness were made available for other subsystems to examine, the emotional schema causing the anxiety was identified and its contents extracted into a format which could be fed into other subsystems. Later on, when Richard's co-worker displayed confidence which others approved of, a contradiction-detection mechanism noticed a discrepancy between reality and the prediction that confidence makes you hated, allowing the prediction to be revised.

Under this model, the system which produces moments of introspective awareness is a subsystem like any other in the brain. This means that it will be activated when the right cues trigger it, and its outputs compete with the outputs of other systems submitting content to consciousness. The circumstances under which the system triggers, and its probability of successfully making its contents conscious, are modified by reinforcement learning. Just as practicing a skill such as arithmetic eventually causes various subsystems to manipulate the content of consciousness in the right order, practicing a skill which benefits from introspective awareness will cause the subsystem generating introspective awareness to activate more often.

Meditation as a technique for generating moments of introspective awareness

Just as there are different forms and styles of therapy, there are also different forms and styles of meditation. All of them involve introspective awareness to at least some degree, but they differ in what that awareness is then used for.

In the example with Richard, his therapist asked him to imagine being confident and to then bring his awareness to *why* that felt uncomfortable. In contrast, a more behaviorally oriented therapist might not have examined the reason behind the discomfort. Rather, they might have taught Richard to notice his reaction to the discomfort, and then use that as a cue for implementing an opposite reaction. Both kinds of therapists would ask their clients to generate *some* introspective awareness, but aiming that awareness at different kinds of features, and using the awareness to trigger different kinds of strategies. The results would correspondingly be very different.

Likewise, systems of meditation differ in how much introspective awareness they produce, what kinds of features the awareness-producing subsystem is trained to extract, and what that awareness is then used for. For this article, I have chosen to use the example of the system in *The Mind Illuminated* (TMI), as it is clearly explained and explicitly phrased in these terms. (Again, TMI has a more precise distinction between introspective attention and introspective awareness, which I am eliding for the sake of simplicity.)

In TMI's system, as in many others, you start with trying to keep your attention on your breath. In terms of our model, this means that you want to keep sensory outputs corresponding to your breath as the main thing in your consciousness.

The problem with this goal is that there is no subsystem which can just unilaterally decide what to maintain as the center of attention. At any given moment, many different subsystems are competing to make their content conscious. So one system might have the intention to follow the breath, and you do it for a while, but then a planning system kicks in with its intention to think about dinner. Such planning has tended to feel rewarding, so it wins out and the intent to meditate is forgotten until five minutes later, when you decide what you want for dinner and then suddenly remember the thing about following your breath.

TMI calls this *mind-wandering from forgetting*, and the first step of practice is just to notice it whenever it happens, congratulate yourself for having noticed it, and then return to the breath. Being able to notice forgetting requires having a moment of introspective awareness which points out the fact that you had not been following your breath. When you take satisfaction in having noticed this, your awareness-producing subsystem gets assigned a reward and becomes slightly more likely to activate in the future. "Have I remembered to follow my breath or not?" acts a feedback mechanism that you can explicitly train on.

As the awareness-producing system starts to activate more often and ping you if you have forgotten to meditate, periods of mind-wandering grow shorter.

Now, even if you stop getting entirely lost in thought, you still have *distraction:* content from other subsystems that is in consciousness together with the sensations of the breath and the intention to focus on the breath. For example, you might be having stray thoughts, hearing sounds from your environment, and experiencing sensations from your body.

To more exclusively focus on the breath, you are instructed to maintain the intent to both attend to it and also to be aware of any distractions. The subsystems which

output mental content can, and normally do, operate independently of each other. This means that the following may happen:

Subsystem 1: I'm meditating well!

Subsystem 2: Hmm, what's that smell.

Subsystem 1: I'm meditating well! No distractions.

Subsystem 2: Smells kinda like cookies.

Subsystem 2: Mmm, cookies.

Subsystem 1: Continuing to meditate well!

Subsystem 2: Say, what's for dinner?

That is, a system which tracks the breath can continue to repeatedly find the breath, and report that your meditation is proceeding well and with no distractions… all the while the content of your consciousness continues to alternate with distracted thoughts, which the breath-tracking subsystem is failing to notice (because it is tracking the breath, not the presence of other thoughts). Worse, since you may find it rewarding to just *think that you are meditating well*, that *thought* may start to become rewarded, and you may find yourself just *thinking* that you are meditating well… even as that thought has become self-sustaining and no longer connected to whether you are following the breath or not!

There are all kinds of subtle traps like this, and reducing the amount of distraction requires you to first have better awareness *of* the distraction. This means more moments of introspective awareness which are tracking what's *actually* happening in your mind:

Subsystem 1: I'm meditating well!

Subsystem 2: Hmm, what's that smell.

Subsystem 1: I'm meditating well! No distractions.

Subsystem 2: Smells kinda like cookies.

Subsystem 2: Mmm, cookies.

A*wareness subsystem:* Wait, one train of thought keeps saying that it's meditating well, but another is totally getting into the thought of food.

Subsystem 1: Oh. Better refocus that attention on the breath, and spend less time thinking about the *concept* of following the breath.

This kind of a process also teaches you to pay attention to patterns of cause and effect in your mind. In this example, the smell of cookies caused you to think of cookies, which in turn made you think of dinner, which could have ultimately led to forgetting and mind-wandering.

Catching the train of thought after "mmm, cookies" meant that three "processing steps" had passed before you noticed it. If you practice tracing back trains of thought in your mind, you seem to teach your awareness-system to collect and store data

from a longer period, even when it is not actively outputting it. This means that at the “mmm, cookies” stage, you can query your awareness to get a trace of the immediately preceding thought chain.

You notice that you started to get distracted starting from the smell of the cookie and can then use this as further input to your awareness system. You are essentially taking the re-presented smell of the cookie which the system output, and feeding it back in, asking it to pay more attention to detecting “things like this”. The next time that you notice a smell, your introspective awareness may flag it right away, letting you catch the distraction at the very first stage and before it turns into an extended train of thought.

Note that there is nothing particularly mysterious or unusual about any of this. You are employing essentially the same process used in learning *any* skill. In learning to ride a bike, for example, attempting to keep the bike balanced involves adjusting your movements in response to feedback. When you do so, your brain becomes better at detecting things like “tilting towards the right” in the sense data, increasing your ability to apply the right correction. After you have learned to identify tilting-a-lot-but-not-quite-falling, your brain learns to backtrace to the preceding state of tilting-a-little-less, and apply the right correction there. Once its precision has been honed to identify that state, you can further detect an even subtler tilt, until you automatically apply the right corrections to keep you balanced.

Essentially the kind of a learning algorithm is being applied here. Increased sensory precision leads to improvements in skill which allow for increased sensory precision. (See also this article, which goes into more detail about TMI as a form of deliberate practice.)

Uses for moments of introspective awareness

I should again emphasize that the preceding explanation is only looking at one particular meditation system. There are other systems which work very differently, but they all use or develop introspective awareness to some extent. For example:

- **In Shinzen Young's formulation of “do nothing” practice,** you have just two basic instructions: *let whatever happens, happen* and *when you notice an intention to control your attention, drop that intention.* This trains introspective awareness to notice when one is trying to control their attention… but it is also a very different system, since maintaining an intention to notice when that happens would also be an attempt to control attention! Thus, one is instructed to drop intentions if one spontaneously notices them, but not to actively look for them.
- **In noting practice,** you are trying to consciously name or notice everything that happens in your consciousness. Introspective awareness is trained to very rapidly distinguish between everything that happens, but is not trained to maintain attention on any particular thing.
- **In visualization practice,** you might create a visual image in your mind, then use introspective awareness to examine the mental object that you’ve created and compare it to what a real image would look like. This gives the subsystem creating the visualization feedback, and helps slowly develop a more realistic image.

Going back to TMI-style introspective awareness, once you get it trained up, you can use it for various purposes. In particular, once you learn to maintain it during your daily life - and not just on the meditation couch - it will bring up more assumptions in your various schemas and mental models. Think of Richard paying attention to the assumptions behind his unwanted reactions and making them explicit, but as something that happens on a regular basis as the reactions come up.

Romeo Stevens described what he called “the core loop of Buddhism”:

> So, what is the core loop?
>
> It's basically cognitive behavioral therapy, supercharged with a mental state more intense than most pharmaceuticals.
>
> There are two categories of practice, one for cultivating the useful mental state, the other uses that mental state to investigate the causal linkages between various parts of your perception (physical sensations, emotional tones, and mental reactions) which leads to clearing out of old linkages that weren't constructed well.
>
> You have physical sensations in the course of life. Your nervous system reacts to these sensations with high or low valence (positive, negative, neutral) and arousal (sympathetic and parasympathetic nervous system activation), your mind reacts to these now-emotion-laden sensations with activity (mental image, mental talk) out of which you then build stories to make sense of your situation.
>
> The key insight that drives everything is the knowledge (and later, direct experience) that this system isn't wired up efficiently. Importantly: I don't mean this in a normative way. Like you should wire it the way I say just because, but in the 'this type of circuit only needs 20 nand gates, why are there 60 and why is it shunting excess voltage into the anger circuits over there that have nothing to do with this computation?' way. Regardless of possible arguments over an ultimately 'correct' way to wire everything, there are very low hanging fruit in terms of improvements that will help you effectively pursue *any* other goal you set your mind to.

Again, we saw an example of this with Richard. He had experienced his father as acting confident and as causing suffering to Richard and others; sensations which his mind has classified as negative. In order to avoid them, a model (story) was constructed saying that confidence is horrible, and behaviors (e.g. negative self-talk) were created to avoid appearing horrible.

Now, this caused problems down the line, making him motivated to try to appear more confident… meaning that there was now a mechanism in his brain trying to prevent him from appearing confident, and another which considered this a problem and tried to make him more confident, in opposition to the first system. See what Romeo means when talking about circuits that only need 20 gates but are implemented using 60?

The article “tune your motor cortex” makes the following claims about muscle movement:

> Your motor cortex automatically learns to execute complex movements by putting together simpler ones, all the way down to control of individual muscles.

> Because the process of learning happens organically, the resulting architecture of neural connections (you can think of them as "hidden layers" in machine learning terms) is not always perfectly suited to the task.
>
> Some local optima of those neural configurations are hard to get out of, and constantly reinforced by using them.
>
> There is some pressure for muscle control to be efficient, and the motor cortex is doing a "good enough" job at it, but tends to stop a fair bit from perfection.
>
> By repeating certain movements and positions over and over again (e.g. during sitting work), we involuntarily strengthen connections between movements and muscles that don't make much sense lumped together.
>
> E.g. control of shoulders might become spuriously wired together with control of thighs (both are often tense during sitting).

There are various mental motions which are learned in basically the same way as physical motions are:

- You learn to calculate 12*13 by a technique such as first multiplying 10*13, keeping the result in your memory, calculating 2*13, and then adding the intermediate results together.
- You learn that a particular memory makes you feel slightly unpleasant, and that flinching away from anything that would remind you of it takes the pain away.
- You learn that this also works on uncomfortable chores, teaching you to keep pushing the thought of them away.
- You learn that your father's behavior is painful to you, and that any confidence reminds you of that, so you learn negative self-talk which blocks you from acting confident.
- You learn that saying "no" to people reminds you of being punished for saying "no" to your parents, but that saying "yes" too often means that you are constantly fulfilling promises to other people - so you learn to avoid situations where you would be asked anything.
- You learn that there's something you can do in your mind to stop feeling upset, so you start ignoring your emotions and any information they might have.
- You learn that if you feel bad about not getting the respect you want, thinking "if only I was good enough at persuasion, I would get what I want" gives you a sense of control - even though this pattern also makes you feel personally at fault when you *don't* get what you want.
- You learn that it's rewarding to punish people who have wronged, so you *always* want to punish someone when something goes wrong - even if there is nobody but reality to punish.
- You learn that it feels good to mentally punish someone who is munching too loud, but actually complaining about it would feel petty, and you've learned that pettiness is frowned upon. So you also learn to block the impulse to say anything out loud, but continue to get increasingly angry about the sound, causing an escalating circle of both the annoyance and the blocking ramping up in intensity.

As with physical movements, these can form local optima that are hard to get out of. Many of them are learned in childhood, when your understanding of the world is limited. But new behaviors continue to build on top of them, so you will eventually end up with a system which could use a lot of optimization.

If you have more introspective awareness of the exact processes that are happening in your mind, you can make more implicit assumptions conscious, causing your brain's built-in contradiction detector to notice when they contradict your later learning. Also, getting more feedback about what exactly is happening in your mind allows you to notice more wasted motion in general.

One particular effect is that, as *Unlocking the Emotional Brain* notes, the mind often makes trade-offs where it causes itself some minor suffering in order to avoid a perceived greater suffering. For example, someone may feel guilt in order to motivate themselves, or experience self-doubt to avoid appearing too confident. By employing greater introspective awareness, one may find ways to achieve their goals *without* needing to experience any suffering in order to do so.

Of course, Buddhist meditation is not the only way to achieve this. Various therapies and techniques such as Focusing, Internal Family Systems, Internal Double Crux, and so on, are also methods which use introspective awareness to reveal and refactor various assumptions. Increased introspective awareness from meditation tends to also boost the effectiveness of related techniques, as well as reveal more situations where they can be employed.

If introspective awareness is so great, why don't we have it naturally?

As with anything, there are tradeoffs involved. Having more introspective awareness can help fix a lot of issues… but it also comes with risks, which I assume is the reason why we have *not* evolved to have a lot of it all the time.

First, it's worth noting that even for experienced meditators, intense emotional reactions tend to shut down introspective awareness. If one of the functions of e.g. fear and anxiety is to cause a rapid response, then excessive amounts of introspective awareness would slow down that response by reducing cognitive fusion. Many emotions seem to inhibit many competing processes from accessing consciousness, so that you can deal with the situation at hand.

Another consideration involves traumatic memories. In the beginning of the article, I suggested that anxiety is a special kind of mental object which activates particular behaviors. In general, different emotional states have specific kinds of behaviors and activities associated with them - meaning that if you have some memories which are *really* painful, they can become overwhelming, making it necessary to block them in order to carry on with your normal life. Meditation can be helpful for working through your trauma, but it can also bring it up before you are ready for it, to the point of requiring professional psychotherapy to get through. If you are better at noticing all kinds of subtle details in your mind, it also becomes easier to notice anything that would remind you of things you don't want to remember. A *decrease* in introspective awareness seems to be a common trauma symptom, as this helps block the unpleasant memories from being too easily triggered.

I have also heard advanced meditators mention that increased introspective awareness makes it difficult to push away pangs of conscience that they would otherwise have ignored, causing practical problems. For example, people have said that they are no longer able to eat animal products or tell white lies.

On the other hand, extended concentration practice can also make it easier to *block* things which you would be better off not blocking.

So far, this article has mostly focused on using introspective awareness to notice the *content* of your thoughts. But you can also use it to notice the *structure* of the higher-level processes generating your thoughts. Part of how you develop concentration ability is by maintaining introspective awareness of the fact that being able to concentrate on just one thing feels more pleasant than having your attention jump between many different things. This can give you an improved ability to choose what you are concentrating on… but also to selectively exclude anything unpleasant from your mind.

For example, there was an occasion when I needed to do some work, but also had intense anxiety about not wanting to; intense enough that it would normally have made it impossible for me to focus on it. So then I *tried* to work, and let my introspective awareness observe the feeling of head-splitting agony from my attention alternating between the work and the desire not to… and to also notice that whenever my attention was on the work, I felt temporarily better.

After a while of this, the anxiety started to get excluded from my consciousness, until it suddenly dropped away completely - as if some deeper process had judged it useless and revoked its access to consciousness. And while this allowed me to do the work that I needed to, it also felt internally violent, and like it would be too easy to repress any unpleasant thoughts using it. I still use this kind of technique on occasion when I need to concentrate on something, but I try to be cautious about it.

The negative side of being able to get better feedback about your mental processes, is that you can also get better feedback on exactly how pleasant wireheading feels. If you like to imagine pleasant things, you can get better and better at imagining pleasant things, and excluding any worries about it from your consciousness. Meditation teacher Daniel Ingram warns:

> Strong insight and concentration practice, even when that practice wasn't dedicated to the powers, can make people go temporarily or permanently (or for the rest of that lifetime) psychotic. The more the practice involves creating experiences that diverge significantly from what I will crudely term "consensus reality", and the longer one engages in these practices, the more likely prolonged difficulties are. It is of note that a significant number of the primary propagators of the Western magickal traditions became moderately nuts towards the ends of their lives.
>
> As one Burmese man said to Kenneth, "My brother does concentration practice. You know, sometimes they go a little mad!" He was talking about what can sometimes happen when people get into the powers. [...]
>
> I remember a letter from a friend who was on a long retreat in Burma and was supposed to be doing insight practices but had slipped into playing with these sorts of experiences. He was now fascinated by his ability to see spirit animals and other supernormal beings and was having regular conversations with some sort of low-level god that kept telling him that he was making excellent progress in his insight practice—that is, exactly what he wanted to hear. However, the fact that he was having stable visionary experiences and was buying into their content made it abundantly clear that he wasn't doing insight practices at all, but was lost in and being fooled by these.

Now, it should be pointed out that “being able to exclude anything unpleasant from your consciousness” is only going to be a worry for advanced practitioners who spend a lot of time on the kind of practice that inclines you towards these kinds of risks. Before you get to the point of something like this being a risk, you will get to resolve a lot of internal conflicts and old issues first.

Here is Culadasa, the author of *The Mind Illuminated,* being interviewed about this kind of a “first you resolve a lot of issues, but then you can get the ability to push down the rest” dynamic:

> Michael Taft: … and you’re using the meditation practice to help work with your stuff. But what about the other case that we both know of where people have reached very high levels of meditative capacity, they’ve got a lot of insight, maybe they’re at some level of awakening, and they seem to have, in a way, missed a whole pocket of material, or several pockets of material. It’s like they think they’re doing fine, but maybe everyone around them is aware that they’ve got these behavior patterns that do not seem awake at all. And yet the meditation has somehow missed that.
>
> Culadasa: Yes, yes. […] ... there seems to be a certain level of the stuff that we’re talking about that it’s necessary to deal with to achieve awakening, but it’s sort of a minimal level. […] What I think that is indicative of is that if that hasn’t been sufficiently dealt with earlier, it has to get dealt with in one way or another at that point. That doesn’t necessarily mean that it’s going to get resolved; it may just get reburied a little more deeply.
>
> Michael Taft: Pushed out of the way.
>
> Culadasa: Yeah, pushed out of the way, or bypassed in some way. That allows a person to go ahead and [progress] and it’s unrealistic to think that everything has been resolved. […] a lot of the things that change […] actually help to push these things aside, to bypass them in one way or another, whereas before somebody has [made as much progress] these would have been sufficiently problematic in their life that, in one way or another, they would be aware of them, whether or not they did anything about them or were at a place of just taking for granted that I have these, quote, “personality characteristics” that are a bit difficult.

I used to be very enthusiastic about TMI’s meditation system. I still consider it important and useful to make progress on, but am slightly more guarded after some of my own experiences, hearing about the experience of a friend who reached a high level in it, reading some critiques of its tendency to emphasize awareness of positive experiences [1 2], and considering both the interview quoted above and Culadasa’s subsequent actions. (That said, the focus on positive experiences can be a useful counterbalance for people who start off with an overall negative stance towards life.)

I continue to practice it, and would generally find it safe until you get to around the sixth or so of its ten stages, at which point I would suggest starting to exercise some caution. Off the couch, I mostly don’t do much concentration practice (except in a context where I would need to concentrate anyway). Rather I try to focus my introspective awareness towards just observing my mind without actively interfering with it, Internal Family Systems -style practice, and other activities that do not seem to risk excluding too much unpleasant material.

Finally, developing too much awareness into your mind may cause you to start noticing contradictions between how you thought it worked, and how it actually works.

I suspect that a part of how our brains have evolved to operate, *relies on* those differences going unnoticed. This gets us to the topic of enlightenment, which I have not yet discussed, but will do in my next post.

Thanks to Maija Haavisto, Lumi Pakkanen and Romeo Stevens for comments on an earlier draft.

Mental Mountains

I.

Kaj Sotala has an outstanding review of *Unlocking The Emotional Brain*; I read the book, and Kaj's review is better.

He begins:

> UtEB's premise is that much if not most of our behavior is driven by emotional learning. Intense emotions generate unconscious predictive models of how the world functions and what caused those emotions to occur. The brain then uses those models to guide our future behavior. Emotional issues and seemingly irrational behaviors are generated from implicit world-models (schemas) which have been formed in response to various external challenges. Each schema contains memories relating to times when the challenge has been encountered and mental structures describing both the problem and a solution to it.

So in one of the book's example cases, a man named Richard sought help for trouble speaking up at work. He would have good ideas during meetings, but felt inexplicably afraid to voice them. During therapy, he described his narcissistic father, who was always mouthing off about everything. Everyone hated his father for being a fool who wouldn't shut up. The therapist conjectured that young Richard observed this and formed a predictive model, something like "talking makes people hate you". This was overly general: talking only makes people hate you if you talk incessantly about really stupid things. But when you're a kid you don't have much data, so you end up generalizing a lot from the few examples you have.

When Richard started therapy, he didn't consciously understand any of this. He just felt emotions (anxiety) at the thought of voicing his opinion. The predictive model output the anxiety, using reasoning like "if you talk, people will hate you, and the prospect of being hated should make you anxious – therefore, anxiety", but not any of the intermediate steps. The therapist helped Richard tease out the underlying model, and at the end of the session Richard agreed that his symptoms were related to his experience of his father. But knowing this changed nothing; Richard felt as anxious as ever.

Predictions like "speaking up leads to being hated" are special kinds of emotional memory. You can rationally understand that the prediction is no longer useful, but that doesn't really help; the emotional memory is still there, guiding your unconscious predictions. What should the therapist do?

Here *UtEB* dives into the science on memory reconsolidation.

Scientists have known for a while that giving rats the protein synthesis inhibitor anisomycin prevents them from forming emotional memories. You can usually give a rat noise-phobia by pairing a certain noise with electric shocks, but this doesn't work if the rats are on anisomycin first. Probably this means that some kind of protein synthesis is involved in memory. So far, so plausible.

A 2000 study found that anisomycin could also erase existing phobias in a very specific situation. You had to "activate" the phobia – get the rats thinking about it really hard, maybe by playing the scary noise all the time – and then give them the

anisomycin. This suggested that when the memory got activated, it somehow “came loose”, and the brain needed to do some protein synthesis to put it back together again.

Thus the idea of memory reconsolidation: you form a consolidated memory, but every time you activate it, you need to reconsolidate it. If the reconsolidation fails, you lose the memory, or you get a slightly different memory, or something like that. If you could disrupt emotional memories like “speaking out makes you hated” while they’re still reconsolidating, maybe you could do something about this.

Anisomycin is pretty toxic, so that’s out. Other protein synthesis inhibitors are also toxic – it turns out proteins are kind of important for life – so they’re out too. Electroconvulsive therapy actually seems to work pretty well for this – the shock disrupts protein formation very effectively (and the more I think about this, the more implications it seems to have). But we can’t do ECT on everybody who wants to be able to speak up at work more, so that’s also out. And the simplest solution – activating a memory and then reminding the patient that they don’t rationally believe it’s true – doesn’t seem to help; the emotional brain doesn’t speak Rationalese.

The authors of *UtEB* claim to have found a therapy-based method that works, which goes like this:

First, they tease out the exact predictive model and emotional memory behind the symptom (in Richard’s case, the narrative where his father talked too much and ended up universally hated, and so if Richard talks at all, he too will be universally hated). Then they try to get this as far into conscious awareness as possible (or, if you prefer, have consciousness dig as deep into the emotional schema as possible). They call this “the pro-symptom position” – giving the symptom as much room as possible to state its case without rejecting it. So for example, Richard’s therapist tried to get Richard to explain his unconscious pro-symptom reasoning as convincingly as possible: “My father was really into talking, and everybody hated him. This proves that if I speak up at work, people will hate me too.” She even asked Richard to put this statement on an index card, review it every day, and bask in its compellingness. She asked Richard to imagine getting up to speak, and feeling exactly how anxious it made him, while reviewing to himself that the anxiety felt justified given what happened with his father. The goal was to establish a wide, well-trod road from consciousness to the emotional memory.

Next, they try to find *a lived and felt experience* that contradicts the model. Again, Rationalese doesn’t work; the emotional brain will just ignore it. But it will listen to experiences. For Richard, this was a time when he was at a meeting, had a great idea, but didn’t speak up. A coworker had the same idea, mentioned it, and everyone agreed it was great, and congratulated the other person for having such an amazing idea that would transform their business. Again, there’s this same process of trying to get as much in that moment as possible, bring the relevant feelings back again and again, create as wide and smooth a road from consciousness to the experience as possible.

Finally, the therapist activates the disruptive emotional schema, and before it can reconsolidate, smashes it into the new experience. So Richard’s therapist makes use of the big wide road Richard built that let him fully experience his fear of speaking up, and asks Richard to get into that frame of mind (activate the fear-of-speaking schema). Then she asks him, *while keeping the fear-of-speaking schema in mind*, to remember the contradictory experience (coworker speaks up and is praised). Then the

therapist vividly describes the juxtaposition while Richard tries to hold both in his mind at once.

And then Richard was instantly cured, and never had any problems speaking up at work again. His coworkers all applauded, and became psychotherapists that very day. An eagle named “Psychodynamic Approach” flew into the clinic and perched atop the APA logo and shed a single tear. *Coherence Therapy: Practice Manual And Training Guide* was read several times, and God Himself showed up and enacted PsyD prescribing across the country. All the cognitive-behavioralists died of schizophrenia and were thrown in the lake of fire for all eternity.

This is, after all, a therapy book.

II.

I like *UtEB* because it reframes historical/purposeful accounts of symptoms as aspects of a predictive model. We already know the brain has an unconscious predictive model that it uses to figure out how to respond to various situations and which actions have which consequences. In retrospect, this framing perfectly fits the idea of traumatic experiences having outsized effects. Tack on a bit about how the model is more easily updated in childhood (because you’ve seen fewer other things, so your priors are weaker), and you’ve gone a lot of the way to traditional models of therapy.

But I also like it because it helps me think about the idea of separation/noncoherence in the brain. Richard had his schema about how speaking up makes people hate you. He also had lots of evidence that this wasn’t true, both rationally (his understanding that his symptoms were counterproductive) and experientially (his story about a coworker proposing an idea and being accepted). But the evidence failed to naturally propagate; it didn’t connect to the schema that it should have updated. Only after the therapist forced the connection did the information go through. Again, all of this should have been obvious – of course evidence doesn’t propagate through the brain, I was writing posts ten years ago about how even a person who knows ghosts exist will be afraid to stay in an old supposedly-haunted mansion at night with the lights off. But UtEB’s framework helps snap some of this into place.

UtEB’s brain is a mountainous landscape, with fertile valleys separated by towering peaks. Some memories (or pieces of your predictive model, or whatever) live in each valley. But they can’t talk to each other. The passes are narrow and treacherous. They go on believing their own thing, unconstrained by conclusions reached elsewhere.

Consciousness is a capital city on a wide plain. When it needs the information stored in a particular valley, it sends messengers over the passes. These messengers are good enough, but they carry letters, not weighty tomes. Their bandwidth is atrocious; often they can only convey what the valley-dwellers think, and not why. And if a valley gets something wrong, lapses into heresy, as often as not the messengers can’t bring the kind of information that might change their mind.

Links between the capital and the valleys may be tenuous, but valley-to-valley trade is almost non-existent. You can have two valleys full of people working on the same problem, for years, and they will basically never talk.

Sometimes, when it’s very important, the king can order a road built. The passes get cleared out, high-bandwidth communication to a particular communication becomes possible. If he does this to two valleys at once, then they may even be able to share

notes directly, each passing through the capital to get to each other. But it isn't the norm. You have to really be trying.

This ended out a little more flowery than I expected, but I didn't start thinking this way because it was poetic. I started thinking this way because of this:

Frequent SSC readers will recognize this as from Figure 1 of Friston and Carhart-Harris' REBUS And The Anarchic Brain: Toward A Unified Model Of The Brain Action Of Psychedelics, which I review here. The paper describes it as "the curvature of the free-energy landscape that contains neuronal dynamics. Effectively, this can be thought of as a flattening of local minima, enabling neuronal dynamics to escape their basins of attraction and—when in flat minima—express long-range correlations and desynchronized activity."

Moving back a step: the paper is trying to explain what psychedelics do to the brain. It theorizes that they weaken high-level priors (in this case, you can think of these as the tendency to fit everything to an existing narrative), allowing things to be seen more as they are:

> A corollary of relaxing high-level priors or beliefs under psychedelics is that ascending prediction errors from lower levels of the system (that are ordinarily unable to update beliefs due to the top-down suppressive influence of heavily-weighted priors) can find freer register in conscious experience, by reaching and impressing on higher levels of the hierarchy. In this work, we propose that this straightforward model can account for the full breadth of subjective phenomena associated with the psychedelic experience.

These ascending prediction errors (ie noticing that you're wrong about something) can then correct the high-level priors (ie change the narratives you tell about your life):

> The ideal result of the process of belief relaxation and revision is a recalibration of the relevant beliefs so that they may better align or harmonize with other levels of the system and with bottom-up information—whether originating from within (e.g., via lower-level intrinsic systems and related interoception) or, at lower doses, outside the individual (i.e., via sensory input or extroception). Such functional harmony or realignment may look like a system better able to guide thought and behavior in an open, unguarded way (Watts et al., 2017; Carhart-Harris et al., 2018b).

This makes psychedelics a potent tool for psychotherapy:

> Consistent with the model presented in this work, overweighted high-level priors can be all consuming, exerting excessive influence throughout the mind and brain's (deep) hierarchy. The negative cognitive bias in depression is a good example of this (Beck, 1972), as are fixed delusions in psychosis (Sterzer et al., 2018).25 In this paper, we propose that psychedelics can be therapeutically effective, precisely because they target the high levels of the brain's functional hierarchy, primarily affecting the precision weighting of high-level priors or beliefs. More specifically, we propose that psychedelics dose-dependently relax the precision weighting of high-level priors (instantiated by high-level cortex), and in so doing, open them up to an upsurge of previously suppressed bottom-up signaling (e.g., stemming from limbic circuitry). We further propose that this sensitization of high-level priors means that more information can impress on

> them, potentially inspiring shifts in perspective, felt as insight. One might ask whether relaxation followed by revision of high-level priors or beliefs via psychedelic therapy is easy to see with functional (and anatomic) brain imaging. We presume that it must be detectable, if the right questions are asked in the right way.

Am I imagining this, or are Friston + Carhart-Harris and *Unlocking The Emotional Brain* getting at the same thing?

Both start with a piece of a predictive model (= high-level prior) telling you something that doesn't fit the current situation. Both also assume you have enough evidence to convince a rational person that the high-level prior is wrong, or doesn't apply. But you don't automatically smash the prior and the evidence together and perform an update. In *UtEB*'s model, the update doesn't happen until you forge conscious links to both pieces of information and try to hold them in consciousness at the same time. In F+CH's model, the update doesn't happen until you take psychedelics which make the high-level prior lose some of its convincingness. *UtEB* is trying to laboriously build roads through mountains; F+CH are trying to cast a magic spell that makes the mountains temporarily vanish. Either way, you get communication between areas that couldn't communicate before.

III.

Why would mental mountains exist? If we keep trying to get rid of them, through therapy or psychedelics, or whatever, then why not just avoid them in the first place?

Maybe generalization is just hard (thanks to MC for this idea). Suppose Goofus is mean to you. You learn Goofus is mean; if this is your first social experience, maybe you also learn that the world is mean and people have it out for you. Then one day you meet Gallant, who is nice to you. Hopefully the system generalizes to "Gallant is nice, Goofus is still mean, people in general can go either way".

But suppose one time Gallant is just having a terrible day, and curses at you, and that time he happens to be wearing a red shirt. You don't want to overfit and conclude "Gallant wearing a red shirt is mean, Gallant wearing a blue shirt is nice". You want to conclude "Gallant is generally nice, but sometimes slips and is mean."

But any algorithm that gets too good at resisting the temptation to separate out red-shirt-Gallant and blue-shirt-Gallant risks falling into the opposite failure mode where it doesn't separate out Gallant and Goofus. It would just average them out, and conclude that people (including both Goofus and Gallant) are medium-niceness.

And suppose Gallant has brown eyes, and Goofus green eyes. You don't want your algorithm to overgeneralize to "all brown-eyed people are nice, and all green-eyed people are mean". But suppose the Huns attack you. You *do* want to generalize to "All Huns are dangerous, even though I can keep treating non-Huns as generally safe". And you want to do this as quickly as possible, definitely before you meet any more Huns. And the quicker you are to generalize about Huns, the more likely you are to attribute false significance to Gallant's eye color.

The end result is a predictive model which is a giant mess, made up of constant "This space here generalizes from this example, except this subregion, which generalizes from this other example, except over here, where it doesn't, and definitely don't *ever* try to apply any of those examples over here." Somehow this all works shockingly well. For example, I spent a few years in Japan, and developed a good model for how

to behave in Japanese culture. When I came back to the United States, I effortlessly dropped all of that and went back to having America-appropriate predictions and reflexive actions (except for an embarrassing habit of bowing whenever someone hands me an object, which I still haven't totally eradicated).

In this model, mental mountains are just the context-dependence that tells me not to use my Japanese predictive model in America, and which prevents evidence that makes me update my Japanese model (like "I notice subways are always on time") from contaminating my American model as well. Or which prevent things I learn about Gallant (like "always trust him") from also contaminating my model of Goofus.

There's actually a real-world equivalent of the "red-shirt-Gallant is bad, blue-shirt-Gallant is good" failure mode. It's called "splitting", and you can find it in any psychology textbook. Wikipedia defines it as "the failure in a person's thinking to bring together the dichotomy of both positive and negative qualities of the self and others into a cohesive, realistic whole."

In the classic example, a patient is in a mental hospital. He likes his doctor. He praises the doctor to all the other patients, says he's going to nominate her for an award when he gets out.

Then the doctor offends the patient in some way – maybe refuses one of his requests. All of a sudden, the doctor is abusive, worse than Hitler, worse than Mengele. When he gets out he will report her to the authorities and sue her for everything she owns.

Then the doctor does something right, and it's back to praise and love again.

The patient has failed to integrate his judgments about the doctor into a coherent whole, "doctor who sometimes does good things but other times does bad things". It's as if there's two predictive models, one of Good Doctor and one of Bad Doctor, and even though both of them refer to the same real-world person, the patient can only use one at a time.

Splitting is most common in borderline personality disorder. The DSM criteria for borderline includes splitting (there defined as "a pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation"). They also include things like "markedly and persistently unstable self-image or sense of self", and "affective instability due to a marked reactivity of mood", which seem relevant here too.

Some therapists view borderline as a disorder of integration. Nobody is great at having all their different schemas talk to each other, but borderlines are atrocious at it. Their mountains are so high that even different thoughts about the same doctor can't necessarily talk to each other and coordinate on a coherent position. The capital only has enough messengers to talk to one valley at a time. If tribesmen from the Anger Valley are advising the capital today, the patient becomes truly angry, a kind of anger that utterly refuses to listen to any counterevidence, an anger pure beyond your imagination. If they are happy, they are *purely* happy, and so on.

About 70% of people diagnosed with dissociative identity disorder (previously known as multiple personality disorder) have borderline personality disorder. The numbers are so high that some researchers are not even convinced that these are two different conditions; maybe DID is just one manifestation of borderline, or especially severe borderline. Considering borderline as a failure of integration, this makes sense; DID is total failure of integration. People in the furthest mountain valleys, frustrated by

inability to communicate meaningfully with the capital, secede and set up their own alternative provincial government, pulling nearby valleys into their new coalition. I don't want to overemphasize this; most popular perceptions of DID are overblown, and at least some cases seem to be at least partly iatrogenic. But if you are bad enough at integrating yourself, it seems to be the sort of thing that can happen.

In his review, Kaj relates this to Internal Family Systems, a weird form of therapy where you imagine your feelings as people/entities and have discussions with them. I've always been skeptical of this, because feelings are not, in fact, people/entities, and it's unclear why you should expect them to answer you when you ask them questions. And in my attempts to self-test the therapy, indeed nobody responded to my questions and I was left feeling kind of silly. But Kaj says:

> As many readers know, I have been writing a sequence of posts on multi-agent models of mind. In Building up to an Internal Family Systems model, I suggested that the human mind might contain something like subagents which try to ensure that past catastrophes do not repeat. In subagents, coherence, and akrasia in humans, I suggested that behaviors such as procrastination, indecision, and seemingly inconsistent behavior result from different subagents having disagreements over what to do.
>
> As I already mentioned, my post on integrating disagreeing subagents took the model in the direction of interpreting disagreeing subagents as conflicting beliefs or models within a person's brain. Subagents, trauma and rationality further suggested that the appearance of drastically different personalities within a single person might result from unintegrated memory networks, which resist integration due to various traumatic experiences.
>
> This post has discussed UtEB's model of conflicting emotional schemas in a way which further equates "subagents" with beliefs – in this case, the various schemas seem closely related to what e.g. Internal Family Systems calls "parts". In many situations, it is probably fair to say that this is what subagents are.

This is a model I can get behind. My guess is that in different people, the degree to which mental mountains form a barrier will cause the disconnectedness of valleys to manifest as anything from "multiple personalities", to IFS-findable "subagents", to *UtEB*-style psychiatric symptoms, to "ordinary" beliefs that don't cause overt problems but might not be very consistent with each other.

IV.

This last category forms the crucial problem of rationality.

One can imagine an alien species whose ability to find truth was a simple function of their education and IQ. Everyone who knows the right facts about the economy and is smart enough to put them together will agree on economic policy.

But we don't work that way. Smart, well-educated people believe all kinds of things, even when they should know better. We call these people biased, a catch-all term meaning something that prevents them from having true beliefs they ought to be able to figure out. I believe most people who don't believe in anthropogenic climate change are probably biased. Many of them are very smart. Many of them have read a lot on the subject (empirically, reading more about climate change will usually just make everyone more convinced of their current position, whatever it is). Many of them have enough evidence that they should know better. But they don't.

(again, this is my opinion, sorry to those of you I'm offending. I'm sure you think the same of me. Please bear with me for the space of this example.)

Compare this to Richard, the example patient mentioned above. Richard had enough evidence to realize that companies don't hate everyone who speaks up at meetings. But he still felt, on a deep level, like speaking up at meetings would get him in trouble. The evidence failed to connect to the emotional schema, the part of him that made the real decisions. Is this the same problem as the global warming case? Where there's evidence, but it doesn't connect to people's real feelings?

(maybe not: Richard might be able to say "I know people won't hate me for speaking, but for some reason I can't make myself speak", whereas I've never heard someone say "I know climate change is real, but for some reason I can't make myself vote to prevent it." I'm not sure how seriously to take this discrepancy.)

In Crisis of Faith, Eliezer Yudkowsky writes:

> Many in this world retain beliefs whose flaws a ten-year-old could point out, if that ten-year-old were hearing the beliefs for the first time. These are not subtle errors we're talking about. They would be child's play for an unattached mind to relinquish, if the skepticism of a ten-year-old were applied without evasion…we change our minds less often than we think.
>
> This should scare you down to the marrow of your bones. It means you can be a world-class scientist and conversant with Bayesian mathematics and still fail to reject a belief whose absurdity a fresh-eyed ten-year-old could see. It shows the invincible defensive position which a belief can create for itself, if it has long festered in your mind.
>
> What does it take to defeat an error that has built itself a fortress?

He goes on to describe how hard this is, to discuss the "convulsive, wrenching effort to be rational" that he thinks this requires, the "all-out [war] against yourself". Some of the techniques he mentions explicitly come from psychotherapy, others seem to share a convergent evolution with it.

The authors of *UtEB* stress that all forms of therapy involve their process of reconsolidating emotional memories one way or another, whether they know it or not. Eliezer's work on crisis of faith feels like an *ad hoc* form of epistemic therapy, one with a similar goal.

Here, too, there is a suggestive psychedelic connection. I can't count how many stories I've heard along the lines of "I was in a bad relationship, I kept telling myself that it was okay and making excuses, and then I took LSD and realized that it obviously wasn't, and got out." Certainly many people change religions and politics after a psychedelic experience, though it's hard to tell exactly what part of the psychedelic experience does this, and enough people end up believing various forms of woo that I hesitate to say it's all about getting more rational beliefs. But just going off anecdote, this sometimes works.

Rationalists wasted years worrying about various named biases, like the conjunction fallacy or the planning fallacy. But most of the problems we really care about aren't any of those. They're more like whatever makes the global warming skeptic fail to connect with all the evidence for global warming.

If the model in *Unlocking The Emotional Brain* is accurate, it offers a starting point for understanding this kind of bias, and maybe for figuring out ways to counteract it.

AlphaStar: Impressive for RL progress, not for AGI progress

DeepMind released their AlphaStar paper a few days ago, having reached Grandmaster level at the partial-information real-time strategy game StarCraft II over the summer.

This is very impressive, and yet less impressive than it sounds. I used to watch a lot of StarCraft II (I stopped interacting with Blizzard recently because of how they rolled over for China), and over the summer there were many breakdowns of AlphaStar games once players figured out how to identify the accounts.

The impressive part is getting reinforcement learning to work at all in such a vast state space- that took breakthroughs beyond what was necessary to solve Go and beat Atari games. AlphaStar had to have a rich enough set of potential concepts (in the sense that e.g. a convolutional net ends up having concepts of different textures) that it could learn a concept like "construct building P" or "attack unit Q" or "stay out of the range of unit R" rather than just "select spot S and enter key T". This is new and worth celebrating.

The overhyped part is that AlphaStar doesn't really do the "strategy" part of real-time strategy. Each race has a few solid builds that it executes at GM level, and the unit control is fantastic, but the replays don't look creative or even especially reactive to opponent strategies.

That's because there's no representation of causal thinking - "if I did X then they could do Y, so I'd better do X' instead". Instead there are many agents evolving together, and if there's an agent evolving to try Y then the agents doing X will be replaced with agents that do X'. But to explore as much as humans do of the game tree of viable strategies, this approach could take an amount of computing resources that not even today's DeepMind could afford.

(This lack of causal reasoning especially shows up in building placement, where the consequences of locating any one building here or there are minor, but the consequences of your overall SimCity are major for how your units and your opponents' units would fare if they attacked you. In one comical case, AlphaStar had surrounded the units it was building with its own factories so that they couldn't get out to reach the rest of the map. Rather than lifting the buildings to let the units out, which is possible for Terran, it destroyed one building and then immediately began rebuilding it before it could move the units out!)

This means that, first, AlphaStar just doesn't have a decent response to strategies that it didn't evolve, and secondly, it doesn't do very well at building up a reactive decision tree of strategies (if I scout this, I do that). The latter kind of play is unfortunately very necessary for playing Zerg at a high level, so the internal meta has just collapsed into one where its Zerg agents predictably rush out early attacks that are easy to defend if expected. This has the flow-through effect that its Terran and Protoss are weaker against human Zerg than against other races, because they've never practiced against a solid Zerg that plays for the late game.

The end result cleaned up against weak players, performed well against good players, but practically never took a game against the top few players. I think that DeepMind

realized they'd need another breakthrough to do what they did to Go, and decided to throw in the towel while making it look like they were claiming victory. (Key quote: "Prof Silver said the lab 'may rest at this point', rather than try to get AlphaStar to the level of the very elite players.")

Finally, RL practitioners have known that genuine causal reasoning could never be achieved via known RL architectures- you'd only ever get something that could execute the same policy as an agent that had reasoned that way, via a very expensive process of evolving away from dominated strategies at each step down the tree of move and countermove. It's the biggest known unknown on the way to AGI.

The Curse Of The Counterfactual

The Introduction

The Curse of the Counterfactual is a side-effect of the way our brains process is-ought distinctions. It causes our brains to compare our past, present, and future to various counterfactual imaginings, and then blame and punish ourselves for the difference between reality, and whatever we just made up to replace it.

Seen from the outside, this process manifests itself as stress, anxiety, procrastination, perfectionism, creative blocks, loss of motivation, inability to let go of the past, constant starting and stopping on one goal or frequent switching between goals, low self-esteem and many other things. From the *inside*, however, these counterfactuals can feel more real to us than *reality itself*, which can make it difficult to even notice it's happening, let alone being able to stop it.

Unfortunately, even though each specific *instance* of the curse can be defused using relatively simple techniques, we can't just remove the parts of our brain that generate new instances of the problem. Which means that you can't sidestep the Curse by imagining yet another counterfactual world: one in which you believe you *ought* to be able to avoid falling into its trap, just by being smarter or more virtuous!

Using examples derived from my client work, this article will show how the Curse operates, and the bare bones of some approaches to rectifying it, with links to further learning materials. (Case descriptions are anonymized and/or composites; i.e., the names are not real, and identifying details have been changed.)

The Disclaimer

To avoid confusion between object-level advice, and the meta-level issue of "how our moral judgment frames interfere with rational thinking", I have intentionally **omitted** any description of how the fictionalized or composite clients *actually* solved the real-life problems implied by their stories. The examples in this article do not promote or recommend any specific *object-level* solutions for even those clients' actual specific problems, let alone universal advice for people in similar situations.

So, if you have the impression that I am recommending, for example, specific ways to deal with career or relationship issues, you are extrapolating something that is *not actually here*: this article is strictly about how the interaction between counterfactuals and moral judgment interferes with our practical thinking processes, not about what *conclusions* people draw once their ability to think practically has been restored.

The Stories

The Wish

Carlos is telling me about his childhood. His father was very strict, imposing cruel and sadistic punishments for the most minor offences. Years ago, the punishments stopped, but Carlos is still upset. His father should not have done those things, he says. "He should have loved me more."

“Is that true?” I ask. “If you compare the statement ‘He should have loved me more’, to what *actually happened*, what do you feel?”

Carlos is hesitant, confused.

I explain further. “Is the truth that he *should* have loved you more? Or is it only true that you *wish* he loved you more?”

Try these two statements on for size, I tell him. How do they feel? Which one is better? Which one is *true*?

- My father *should* have loved me more
- I *wish* my father had loved me more

The first feels angry. Resentful. He feels like a victim, helpless. There is nothing he can do.

The second one, when he tries it, feels different. It is sad, because what he wishes did not come to pass. At the same time, it is wistful, because he is experiencing a glimpse of what it would be like, if his father *had* loved him more. And then the feeling of sadness passes. There is grief, but then it’s *over*.

When he looks back on the past, it’s now just a memory. He still *wishes* that things were different, still feels wistful... but he’s no longer a victim, at least in that particular way.

The Principle

Sara is telling me about a professional conference she recently attended. As part of a group exercise, she tried hard to persuade her group to adopt her plan for their presentation, and was met with dismissal and obstruction.

She is angry at herself for not being better at persuading them. She should’ve been less stubborn, she says. Should have listened more, tried to understand their points of view beforehand, so she could be more persuasive later. If they understood what she had to offer, she thinks, they would have used her ideas.

I explain to Sara that she is suffering from the Curse of the Counterfactual: the brain’s tendency to attach **moral weight** to the things that we imagine *could* have gone differently, or how we believe things *ought* to be.

She is suffering from a more complex version of the Curse than Carlos, but the result is similar. She feels angry at herself, not her father. And she feels at fault for her perceived failings, because her brain is *literally* punishing her for what she failed to do, with guilt and self-directed anger.

“Compare your experience of the event with what you think *should* have happened. Is it true that you *should* have been less stubborn? Or do you only *wish* you had been less stubborn?”

Sara fights the question more than Carlos. Less experienced in emotional reflection, she retreats to a logical argument, saying that it’s *definitely* true she should have been less stubborn, because that would have produced better results.

Her brain, I explain, is in a loop. On the one hand, she knows the facts of what *actually* happened. She admits that she did not *actually* do any of the things she thinks she should have. But her brain persists in arguing that **reality is wrong**. Her brain is telling her it should not have happened the way it did, and that (in effect), the fact that it *did* happen that way must be **punished**.

I explain to Sara that our brain has special machinery devoted to punishing. It makes us feel anger or disgust when we perceive our standards – or our tribe's standards – being violated. And it generally doesn't stop, until or unless the violator is sufficiently punished, or repents.

But *reality* cannot be punished. And it certainly can't ever repent! So every time she thinks of what *did* happen, her brain keeps on punishing her, telling her that it should *not* have happened the way that it did.

"Have you ever heard the phrase, 'it's the principle of the thing'?" I ask. "People go to ridiculous lengths when a principle is at stake, because our brains want to make it costly for others to cross us. The problem is, when we apply 'principles' to reality, the only person who gets punished is *us*."

The Punishment Myth

Ingvar is having trouble getting his work done. He believes he *should* be able to knock it out in an afternoon, but he doesn't. He surfs the internet, feeling guilty the entire time, because he *should* be working.

Ingvar is experienced at Focusing and IFS, so he has better access to his felt sense than Sara or Carlos, and we rapidly dismantle the moral belief that he is a bad person whenever he is not working. Afterwards, we test his prior thought that he should be able to get his work done in a certain amount of time, and he spontaneously begins talking about strategies for getting it done more easily and quickly. He no longer feels stuck about doing the work, the way he did before.

I explain to him that this is because the brain has different machinery for different types of motivation. Our moral judgment system does not motivate us to actually *accomplish* anything. All it can do is motivate us to punish or protest, to rage and repent. "After all," I say. "Punishment doesn't actually change your behavior in any meaningful way. When you weren't doing your work, you punished yourself constantly while surfing the internet. But you never actually *stopped*."

"In fact, while you were punishing yourself, you got to feel *good* about yourself, because punishing meant you *cared*. You weren't some bad person who wouldn't even feel *guilty* about not working. So your self-punishment actually gave you a moral license to continue as you were."

"Yeah," Ingvar says. "You're right. I felt guilty, but also, *better*. If I hadn't been punishing myself it would have been worse, because I would've felt like a *bad person*."

"Exactly. Exercising moral judgment makes us feel good and righteous, because our brain wants to reward us for punishing violators. But because it works this way, it hijacks our actual motivation to *accomplish* anything. The act of punishing *feels* like we're accomplishing something, so we don't feel like doing anything else.

“In addition, while all that is going on, our brain’s creative, problem-solving modules are idle. That’s why you were stuck before. The ideas you’re coming up with now, for *how* to do the work, are not things you thought of before. Some of them, I thought of when you first told me about your problem. But I didn’t *mention* any of them, because from where you were before, you would have said, “yeah, but…” to them. Am I right?”

Ingvar admits that this, too, is true. The very same ideas he is coming up with now to get his work done, would have felt irrelevant, useless, or even insulting had someone suggested them to him thirty minutes ago… because they wouldn’t have helped him *punish* anybody!

The “Nice Guy” Paradigm

I’m explaining the same thing to Sara. She’s protesting that if she *doesn’t* think she should be less stubborn, then how will she ever change it?

“What we *want* and what we think we *should* are two different things. If it truly would be better to be less stubborn, if that’s something you *want*, then not having the ‘should’ actually makes it *easier*.

“But what your brain is doing right now is not *wanting* to be better. Rather, your brain is trying to cancel out a **loss**.

“Right now, you are imagining a way that you would prefer things to have happened at the conference. But the fact that it didn’t happen that way is painful, because the way things actually went is not as good as what you imagined or hoped for. But if you say to yourself that you *should* have acted differently, then it allows your brain to preserve *hope*.

“If you believe you *should* have acted differently, then you can continue to believe that they *would* have accepted your ideas, if only you had been better at convincing them. It’s like holding on to a bad investment and not selling it, so you don’t have to acknowledge the loss in your mental bookkeeping.”

The specific variant of the Counterfactual Curse that Sara is experiencing is the “Nice Guy Paradigm”. (Which, despite the name, is not actually gender-specific; it’s actually from a book called “No More Mr. Nice Guy”, about becoming assertive instead of people-pleasing.)

The Nice Guy Paradigm is any belief of the form, “If I were X enough, then [other people / reality / I] would [do / be / have] Y.” (In the book’s original formulation, this was expressed more concretely as, “IF I can hide my flaws and become what I think others want me to be, THEN I will be loved, get my needs met, and have a problem-free life.”)

In Sara’s case, she believes that *if only* she were good enough at persuasion, not being stubborn, etc. then people would understand and accept her ideas (and respect and appreciate her).

The upside of this belief is that it allows her to continue hoping that *someday*, maybe she will be good enough at these things, so *eventually* she will get the respect and acknowledgment she deserves and desires. (Which helps her feel less bad about the fact that *today*, she is not getting those things.)

The downside of this belief, though, is that since what other people do is not 100% within her control, she could be the world's best persuader and *still* get let down sometimes. And because this belief runs backwards as well as forwards, then when other people *don't* acknowledge or respect her, she will still feel that it is *all her fault*. (And it will never cross her mind that some people might just be dicks or just plain *unwilling* to understand or accept her or her ideas, *no matter how good* she or those ideas may be.)

The Bad News

Another client, Victor, is excited. I've just explained the curse of the counterfactual as it relates to his problem. "So I should just stop using 'should' and everything will be better?"

"No, sorry. It doesn't work that way. The part of your brain that ties counterfactual imaginings to moral judgment isn't going to go away by us wishing it would. We can remove the *links* from the activities and situations that *trigger* the "shoulds", and we can *specifically* question the truth of individual "shoulds" to get free of them. But it is not an intellectual exercise. It's an **experiential** one.

"To put it another way, your moral judgment system can be persuaded that it made a mistake about whether to punish *this one thing in particular*, but it cannot be persuaded that *it's a mistake to punish things in general*. (Motivating you to punish things is what that part of your brain *does*, after all; it's not like it can go get another job!)"

I tell him about the time I first found out about the Curse and how to fix individual instances of it, and how I, too, thought that *I* "should" be able to "just stop using shoulds". (And I'm not proud to say it took me *years* to fully realize the inherent meta-contradiction taking place there!)

I tell Victor about a book on the process we've just used to tackle one of his problems, and mention that there's a chapter in it devoted to a session where the issue somebody wants to work on is *this very one*: the fact that they think they *should* be able to fix all their problems without having to individually address each and every "should" they have.

Victor laughs once he sees the "meta" of it, the inherent contradiction that nonetheless took me years to beat into my own skull. "So I should probably work on that first, yeah?"

Probably so, Victor. Probably so.

The Theory

The Bias

Byron Katie has a wonderful term we can use to name an instance of the Curse. She calls it "an argument with reality". Because our brain is arguing that, because it can imagine something *better* than whatever actually happened, then, in some vaguely "moral" sense, that better thing *ought* to have happened instead.

But, since that better thing *didn't* happen, that clearly means reality is **wrong**, and *someone* must therefore be punished.

(Maybe you!)

But reality, no matter how repugnant it may be (morally or otherwise), and no matter how much we want to punish it, is still *reality*.

And as Byron Katie puts it, "When I argue with reality, I *lose*... but only 100% of the time!"

Now, to our moral brains, this statement may *itself* seem morally wrong. "How dare you!" our brains may say. "How dare you imply that we should forgive/accept/approve historical atrocity X!"

How *dare* you tell us to accept the existence of suffering, death, imperfection?

It is important to understand that this is an illusion, a *bias*. When activated, the moral brain acts as though the **only** thing motivating *anyone* is proper punishment and disapproval. It makes us feel that, if we fail to be sufficiently outraged, then *nothing* will ever happen. Justice will *never* be done.

And it does this to us, because, for the good of the tribe – that is to say, the good of our genes! – we must be motivated to not only punish the wrongdoers, we must *also* be motivated to punish the non-punishers.

So when you first consider the possibility of accepting reality, over your moral brain's objections, it will feel like you are arguing for the collapse of civilization, and the abandonment of everything you hold dear.

Do not believe this.

The Difference

You can *want* to end death, disease, and suffering, without rejecting the *reality* of death, disease and suffering.

Moral judgment and preferences are two entirely different and separate things. And when moral judgment is involved, trade-offs become taboo.

When Ingvar was procrastinating, and felt he *should* do his work faster, his brain spent absolutely **zero** time considering *how* he might get it done at all, let alone how he might do it faster.

Why? Because to the moral mind, the *reasons* he is not getting it done *do not matter*. Only punishing the evildoer matters, so even if someone suggested ways he could make things easier, his moral brain rejects them as irrelevant to the *real* problem, which is *clearly* his moral failing. Talking or thinking about problems or solutions isn't really "working", therefore it's further evidence of his failing. And making the work *easier* would be lessening his rightful punishment!

So when moral judgment is involved, actually *reasoning* about things feels **wrong**. Because reasoning might lead to a compromise with the Great Evil: a lessening of punishment or a toleration of non-punishers.

This is only an illusion, albeit a *very persistent one*.

The truth is that, when you switch off moral judgment, *preference* remains. Most of us, given a choice, actually *prefer* that good things happen, that we actually act in ways that are good and kind and righteous, that are not about fighting Evil, but simply making *more* of whatever we actually want to see in the world.

And ironically, we are *more* motivated to actually *produce* these results, when we do so from preference than from outrage. We can be creative, we can plan, or we can even compromise and *adjust* our plans to work with reality as it *is*, rather than as we would prefer it to be.

After all, when we think that something is how the world *should* be, it gives us no real motivation to *change* it. We are motivated instead to *protest* and *punish* the state of the world, or to "speak out" against those we believe responsible... and then *feel* like we just accomplished something by doing so!

And so we end up just like Ingvar, surfing the net and punishing himself, but never *actually* working... nor even choosing *not* to work and to do something more rewarding instead.

The Way Out

The Methods

There are many methods we can use to combat the curse of the counterfactual.

For example, the Litany of Gendlin tells us that admitting to reality *cannot* make it worse, because whatever is happening, we are already enduring it. (It just doesn't *feel* that way, while the mind is still clinging to its counterfactuals, as if it were a corporate executive putting off writing down a bad investment, so as not to affect the shareholders' annual report!)

We can also use the Litany of Tarski, and tell ourselves that if we live in a world where the counterfactual is true, then we need to know that, but conversely, if we live in a world where it is *not* true, then we need just as much to know that, too.

These litanies, however, are more of a reminder that points to a thing, than the actual thing itself. They remind us and prompt us to wrestle with the truth (or our idea of it), but they aren't a substitute for *actually doing so*.

So the primary technique I use and teach for actually engaging with the brain's moral judgment system (and then switching it off), is a variation on The Work of Byron Katie.

The Work is a process that in its simplest form consists of a few questions that, when asked in the right way, can gently lead our brain to notice that 1) our counterfactuals are not reality, 2) thinking they *are* reality is *painful*, and 3) maybe it would feel better if we *didn't* think that way any more. A little ditty describing the process goes, "Judge your neighbor, write it down; ask four questions, turn it around."

The reason it begins with "judge your neighbor" is that the technique was originally created to deal with external moral judgments about what other people should or shouldn't do. (Like, "my father should have loved me more".) The technique is a little easier to use on such judgments, presumably because our moral system is more

oriented towards judging other people than abstract concepts. (So using it on judgments of yourself can be a good bit more challenging if you haven't *first* practiced it in the way it was intended to be used.)

In this article, I am not going to get into much detail on the process, as there are free downloads at Byron Katie's website, and she has two excellent books (*Loving What Is*, and *I Need Your Love: Is That True?*) containing transcript after transcript of people doing the process on a wide variety of beliefs, as well as additional exercises for discovering one's judgments in the first place. Instead, I want to share the unique variations and caveats that I have learned and refined to both make the process itself clearer, and to make it easier to teach to others, especially people who are more systematically-minded and less "woo" than average.

(Note: some of Byron Katie's books discuss sensitive topics including rape, child abuse, war atrocities, and more. In addition, some of this discussion includes having victims question their belief that such things "should not" have happened or the belief it was not their fault. And based on some reviews I've seen online, this is apparently even *more* triggering for some people than hearing about the actual events, once their moral outrage kicks in.)

The Tests

One of the biggest challenges in learning self-help techniques (or rationality techniques, for that matter), is not knowing how something is supposed to *feel from the inside*. We can hear people telling us to believe in ourselves, to let go and accept things, or whatever, but unless we have a way to know what these things are *like*, we cannot know if we're making progress at actually doing them.

For this reason, one of the most important things I do as a mindhacking instructor is to develop *tests* that one can apply to one's experience, to know if a technique is being correctly applied.

For the Work of Byron Katie, there are two primary tests that I use and teach, for the first and fourth questions, to know if you are asking the questions correctly, or actually paying attention to your answers.

The first question of the Work is simply, "Is that true?" But it's not looking for what your *reasoning* says, because in the presence of moral judgment, *all* reasoning is motivated reasoning. (Like Sara arguing that it's true she should've been a certain way, because it would have made things better... because things would have been better if she'd done things a certain way.)

Instead, the *real* question we are asking is something more like, "if you reflect on your experience of what has happened/been happening in *reality*, is it actually consistent with the way you're insisting it's supposed to be?"

And the most important part of that question is not "is it consistent?" but "**if you reflect on your experience**." The thing that actually produces a loosening of your moral judgment is not your reasoning about the facts, but the *process* of inquiring into your *experience* of them, and your inward *reflection* on what that means.

This distinction is why the Work is easier for those who have easy access to their inner experience, a skill honed by Focusing, IFS, and various other therapeutic or self-help modalities. But even though those people have the *ability* to access their inner

experience, that doesn't necessarily mean they will actually do so. When contemplating this question, we are **all** tempted sometimes to deflect, to distract, to deny the very possibility of, rather than actually *investigating*.

Because of this, Work facilitators are trained to reject answers to this question other than "yes", or "no", because from the outside, this is the primary "tell" that lets them know whether you're doing this process correctly. If you are answering with something other than "yes" or "no" – for example, if you begin some kind of explanation or story or justification – they immediately know you aren't reflecting on your experience, but providing reasons *not to*, or creating distractions so you won't *have to*.

Unfortunately, while requiring an answer of "yes or no" keeps a facilitator from being sidetracked by your reasoning and distractions, it doesn't actually fix the problem of "not reflecting on your experience", or help with even knowing whether you're reflecting on your experience to begin with.

The First Test

But after many years of doing and teaching this process, I have noticed that there are certain patterns in the *results* of reflecting on one's inner experience in response to the question "is that true?" Whenever I or my clients do the process correctly, there are actually *three* possible answers, not just two, when you take into account how you *feel*.

If you are correctly reflecting on "is that true?", the *experience* of your answer will be similar to one of these three descriptions:

- [feeling of lightness, release, relief] "Huh... I guess that's *not* true." (e.g. "most people should like me... huh, yeah, no, I guess there's not any reason for that to be true")
- [feeling of heaviness, oppression] "I know it doesn't make any sense, but it still *feels* true" (e.g. "I'm bad for not doing my work... I don't want to be, but it feels like *that's just how it is*.")
- [feeling of longing or regret] "I *wish* it were" (e.g. "I *wish* my father loved me more, but I guess it didn't actually happen")

Without understanding this sorting, people often confuse the experience of *wishing* it were true and it actually *being* true. So they answer "yes, it's true" to the question, because the feeling of wishing it true, is rather similar to the feeling that something *is* true.

However, when compared to the heavy feeling of "I hate that it's true", the sad feeling of "I wish it were true" is a bit different, and once identified, can be handled much more easily. (It's fairly simple, after all, to take the admission that you *wish* something were true, and from there, further admit that this means it's actually *not*.)

Or, if you are feeling like it's a bad-but-true thing *oppressing* you, then you are at least making progress of a different kind. You now know that you have an implicit belief or emotional schema *you don't endorse*; that you simply learned at some point that this thing was a moral standard of your tribe. (And knowing this, you can shift to a process more suited for eliciting and correcting such beliefs.)

Or, you can also use The Work's question 2: "Can I absolutely *know* that it's true?" This question can help to loosen the sense of "rightness", inviting you to consider how

you could *possibly* know with 100% certainty the *actual* truth of an “ought”, rather than an “is”, and whether you could make that distinction in practice. (To use a legal analogy, it’s a bit like asking if there’s any conceivable *doubt*.)

The Second Test

For brevity’s sake, we’ll skip a detailed treatment of the Work’s other questions, jumping straight to the outcome of question 4: “Who would you be without that thought?” That is, what is your inwardly-reflected, *simulated experience* of how you would behave, if you weren’t thinking your “should”?

In my experience, the most common failure mode people have for this question is what I call “happy-ever-aftering”. Instead of allowing their mind to automatically generate a simulation based on the what-if, they try to specifically and *deliberately* envision themselves being a better person...

And then fail to notice their feeling that something is *wrong!*

Because about as often as not, the real, *experiential* answer to “Who would you be if you weren’t thinking X?”, is actually an **objection**, **reservation**, or other form of **argument** from your brain.

The thing you imagine doesn’t feel *real* or realistic. Or worse, you feel like you would be a *bad person* in some way, if you stopped thinking or believing the moral judgment. Or perhaps some *bad consequence* would happen, like maybe everybody would stop caring about their work and then nobody would have any coffee and civilization would fall apart.

These reservations can be subtle, but *ignoring* them will make the process fail. You may briefly feel better, having imagined a different “better world” than before, but will soon be disappointed because the oppressive “should” will return, as strong as ever. (Or perhaps be replaced by the idea that you “should” be the better person you imagined at this step!)

But what a reservation or objection simply means, is that your brain has *another* “should” in effect.

For example, at one point when I began this process, I felt that “I should be doing something” when I was trying to go to sleep. When I got to the part about “who I would be without that thought”, I realized that I would feel *worse*, like a “bad person”. Further inquiry showed that this was because I believed that *not* worrying about doing things meant I “wasn’t taking things seriously enough” – a new level of moral judgment to question the truth of.

If we think of our moral judgments as a belief network, where some beliefs are central (“you should take things seriously – i.e., worry about them”) and others less so (“you should be doing something right now”), then most of the time, we are only *aware* of the non-central ones. In our day to day lives, for example, we may often think things like, “I should have done this by now”, but only rarely do we *explicitly* think things like, “I’m a bad person if I’m not working.”

So when we begin the Work of eliminating these harmful judgments, we will nearly always be starting somewhere *shallow*. Thus, the real value of doing the process isn’t that it will fix the first thought we work on (e.g. “I should be doing something”), but

that it will *lead us* to the deeper thoughts (e.g. “I’m a bad person”), through the *objections* or *reservations* we have about changing the first, shallower thoughts.

Then, once we are aware of those deeper beliefs, we can take steps in turn to change *those*. And finally, once they’re no longer supported by these central “strategic” beliefs (I’m bad/not serious/etc.), the everyday, “tactical” beliefs (I should be doing something) tend to fall away on their own.

And then we can actually *think* about solving our real problems, instead of merely punishing ourselves for not having succeeded yet.

The Conclusion

The Curse of the Counterfactual is a side-effect of the way our brains process is-ought distinctions. It causes our brains to compare our past, present, and future to various counterfactual imaginings, and then blame and punish ourselves for the difference between reality, and whatever we just made up to replace it.

Seen from the outside, this process manifests itself as stress, anxiety, procrastination, perfectionism, creative blocks, loss of motivation, inability to let go of the past, constant starting and stopping on one goal or frequent switching between goals, low self-esteem and many other things. From the *inside*, however, these counterfactuals can feel more real to us than *reality itself*, which can make it difficult to even notice it's happening, let alone being able to stop it.

To counteract and fix this tendency, we can use various techniques (such as the litanies of Gendlin and Tarski, and the Work of Byron Katie). But doing so is inherently effortful, in a way that *cannot* be bypassed by mere understanding. There are, however, skills we can learn that make it easier, and tests we can apply to our inner experience that can help us know if we’re making progress or not.

There is no permanent or universal cure for the Curse, but reflecting on our experience in the right ways can release us from individual, *specific* cases of it. And applied closer to the root or center of our belief networks, it can even produce broader, more dramatic shifts in our behavior and what we think of ourselves.

But that’s a topic for another article, as this post is now almost as long as a short ebook!

The Addendum

Speaking of short ebooks, if you’re interested in *other* bugs in the brain that switch off our problem-solving and creativity subsystems, you may want to grab a free copy of A Minute To Unlimit You, as I am currently soliciting feedback on it.

The specific kind of “stuck” it deals with is the kind where you are under pressure to do something, but all you can think about is why you *can’t* do it, what’s stopping you, how you don’t know what to do or can’t decide, etc., instead of anything actually *helpful*.

So if you have a problem like that, I'd appreciate your (emailed) feedback on the content. (Are the instructions clear? Were you able to apply the technique? What

happened afterwards? Just hit "reply" on the receipt email after your download to answer.) Thanks!

Instant stone (just add water!)

This is a linkpost for https://rootsofprogress.org/instant-stone-just-add-water

Originally posted on The Roots of Progress, January 6, 2018

From the time that humans began to leave their nomadic ways and live in settled societies about ten thousand years ago, we have needed to build structures: to shelter ourselves, to store our goods, to honor the gods.

The easiest way to build is with dirt. Mud, clay, any kind of earth. Pile it up and you have walls. A few walls and a thatched roof, and you have a hut.

Earthen hut with thatched roof in Sudan - Petr Adam Dohnálek / Wikimedia

But earthen construction has many shortcomings. Dirt isn't very strong, so you can't build very high or add multiple stories. It tends to wash away in the rain, so it really only works in hot, dry climates. And it can be burrowed through by intruders—animal or human.

We need something tougher. A material that is hard and strong enough to weather any storm, to build high walls and ceilings, to protect us from the elements and from attackers.

Stone would be ideal. It is tough enough for the job, and rocks are plentiful in nature. But like everything else in nature, we find them in an inconvenient form. Rocks don't come in the shape of houses, let alone temples. We could maybe pile or stack them up, if only we had something to hold them together.

If only we could—bear with me now as I indulge in the wildest fantasy—pour *liquid stone* into molds, to create rocks in any shape we want! Or—as long as I'm dreaming—what if we had a glue that was as strong as stone, to stick smaller rocks together into walls, floors and ceilings?

This miracle, of course, exists. Indeed, it may be the oldest craft known to mankind. You already know it—and you probably think of it as one of the dullest, most boring substances imaginable.

I am here to convince you that it is pure magic and that we should look on it with awe.

It's called cement.

Let's begin at the beginning. Limestone is a soft, light-colored rock with a grainy texture, which fizzes in the presence of acid. Chalk is a form of limestone. What distinguishes limestone and makes it useful is a high calcium content ("calcium" and "chalk" are cognates). Specifically, it is calcium carbonate ($CaCO_3$), the same substance that makes up seashells. In fact, limestone, a sedimentary rock, is often formed from crushed seashells, compressed over eons.

Limestone from a quarry in southern Germany - *Hannes Grobe / Wikimedia*

Limestone can be used for many purposes, including fertilizer and whitewash, but its most important industrial use is in making cement. When it is heated to about 1,000 °C (e.g., in a kiln), it produces a powder called quicklime. Chemically, what's going on is that burning calcium carbonate removes carbon dioxide and leaves calcium oxide ($CaCO_3$ + heat → $CaO + CO_2$).

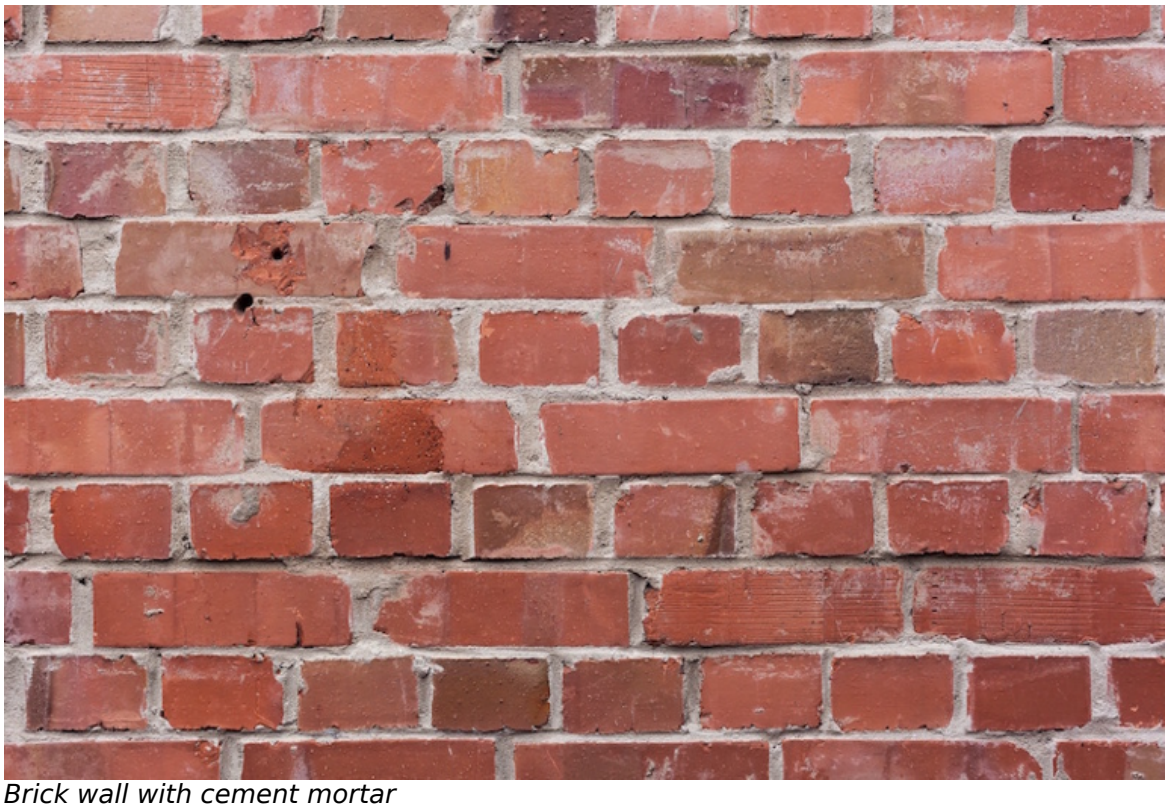
Quicklime is a caustic substance: touching it will burn your skin (hence "quick", meaning active, "alive"). But perhaps its strangest property is that when mixed with water, it reacts, giving off heat—enough to boil the water! The result, called "slaked" or "hydrated" lime, is calcium hydroxide ($CaO + H_2O \rightarrow Ca(OH)_2$ + heat).

Further, if you pour a lime-water slurry into a mold, not too thick, and expose it to the air, a still more amazing thing happens: in a matter of hours, the mixture "sets" and becomes once again as hard as stone. The calcium hydroxide has absorbed CO_2 from the air to return to calcium carbonate ($Ca(OH)_2 + CO_2 \rightarrow CaCO_3 + H_2O$), completing what is known as the "lime cycle".

In other words, by mixing with water and air, this powder—a basic cement—has turned back into rock! If this technology hadn't already existed since before recorded history, it would seem futuristic.

The product of a pure lime cement is too brittle and weak to be very useful (except maybe as a grout). But we can make it stronger by mixing in sand, gravel or pebbles, called "aggregate". Cement, water and sand produce mortar, a glue that can hold together bricks or stones in a masonry wall. Adding gravel or pebbles as well will make concrete, which can be poured into molds to set in place. (The terms "cement" and "concrete" are often

conflated, but technically, cement is the powder from which mortar and concrete are made; concrete is the substance made by adding aggregate and is what constitutes sidewalks, buildings, etc.)



Brick wall with cement mortar

Concrete wall with aggregate visible

This basic technology has been known since prehistoric times: the kilning of limestone is older than pottery, much older than metalworking, and possibly older than *agriculture.* But over the millenia, better formulas for cement have been created, with superior mixtures of ingredients and improved processes.

Pure lime cement needs air to set, so it can't set if poured too thick, or underwater (for instance, on a riverbed to form the base of a column for a bridge). The Romans, who were great users of cement, discovered that adding volcanic ash, called *pozzalana,* to lime would produce a cement that sets even underwater; this is called a "hydraulic cement". They used this "Roman cement" to build everything from aqueducts to the Colosseum. Another common hydraulic cement, called "natural cement", is formed from a mixture of limestone and clay, which sometimes occur together in natural deposits.

Since the mid-1800s, the most widely used cement is a type called Portland cement. Without going into too much detail, this is made through an unintuitive process that involves heating a lime-clay slurry to the point where it fuses together into a hard substance called "clinker". Clinker was originally considered waste material, a ruined product—until it was discovered that grinding it into powder produced a cement that is stronger than Roman or natural cement. (!) Today a wide variety of cements are available on the market, optimized for different conditions.

No matter the formula, however, all cements have one shortcoming: they are very strong under compression, which is the kind of strength needed in a column or wall, but weak under tension, which comes into play, for instance, when a beam buckles under load. The Romans dealt with this problem using arches, which direct forces into compression along the arch. Medieval builders created the pointed Gothic arch, which could stretch even higher than the

round Roman ones, and the flying buttress, which added support to the walls of their tall cathedrals.

Pont du Gard, a Roman aqueduct bridge near Nîmes, France

Gothic window, Church of St. Helen, Lincolnshire, England - Spencer Means / Flickr

But in the twentieth century, a new way of building took over: reinforcing the concrete with steel. Steel, unlike concrete, has high tensile strength, so this “reinforced concrete” is strong under both compression and tension. The reinforcement bars created for this purpose are called “rebar.” Reinforcement allows concrete to be used not only for foundations, walls and columns, but for cantilevered structures such as the decks of Fallingwater.

Fallingwater, by Frank Lloyd Wright - *Mathieu Thouvenin / Flickr*

This is cement. We start with rock, crush and burn it to extract its essence in powdered form, and then reconstitute it at a place and time and *in a shape* of our choosing. Like coffee or pancake mix, it is “instant stone—just add water!” And with it, we make skyscrapers that reach hundreds of stories high, tunnels that go under the English channel and the Swiss Alps, and bridges that stretch a hundred miles.

If that isn’t magic, I don’t know what is.

Sources and further reading: *Concrete Planet: The Strange and Fascinating Story of the World’s Most Common Man-Made Material*, Geology.com, Minerals Education Coalition, Portland Cement Association, and many pages on Wikipedia. Thanks also to Doug Peltz of Mystery Science for helpful conversations.

Could someone please start a bright home lighting company?

This is a linkpost for http://www.lincolnquirk.com/2019/11/26/lumenator.html

Elevator pitch: Bring enough light to simulate daylight into your home and office.

This idea has been shared in Less Wrong circles for a couple years. Yudkowsky wrote Inadequate Equilibria in 2017 where he and his wife invented the idea, and Raemon wrote a playbook in 2018 for how to do it yourself. Now I and at least two other friends are trying to build something similar, and I suspect there's a bigger-than-it-looks market opportunity here because it's one of those things that a lot of people would probably want, if they knew it existed and could experience it. And it's only recently become cheap enough to execute well.

Coelux makes a high-end artificial skylight which certainly looks awesome, but it costs upwards of $30k and also takes a lot of headroom in the ceiling. Can we do better for cheaper?

Brightness from first principles

First let's clear up some definitions:

- Watts is a measure of power consumption, not brightness.
    - "Watt equivalent" brightness is usually listed for LED bulbs, at least for the standard household bulb form factor. You should generally ignore this (instead, just look at the lumens rating), because it is confusing. Normally "watt equivalent" is computed by dividing lumens by 15 or so. (bulb manufacturers like to make LED bulbs that are easy to compare, by having similar brightness to the incandescents they replace, hence "watt equivalent")
- Lumens output is a measurement of an individual bulb, but says nothing about the distribution of those rays of light. For that you want to be doing math to estimate lux.
- "Lux", or "luminous flux", is the measurement of how bright light is on a certain surface (such as a wall or your face). Lux is measured in lumens per square meter. Usually, your end goal when designing lighting is to create a certain amount of lux.
    - Direct sunlight shines 100k lux (source for these on Wikipedia)
    - Full daylight (indirect) is more than 10k lux
    - An overcast day or bright TV studio lighting is 1000 lux
    - Indoor office lighting is typically 500
    - Indoor living room at night might be only 50

Side note: This scale surprises me greatly! We usefully make use of vision with four or more orders of magnitude differences in lux within a single day. Our human vision hardware is doing a lot of work to make the world look reasonable within these vast

differences of amount of light. Regardless, this post is about getting a lot of lux. I hypothesize that lux is associated with both happiness and productivity, and during the "dark season" when we don't get as much lux from the sun, I'm looking to get some from artificial lights.

If you put a single 1000-lumen (66-watt-equivalent) omnidirectional bulb in the center of a spherical room of 2m radius (which approximates a 12' square bedroom), the lux at the radius of the sphere is 50. So now we can get a sense of the scope of the problem. When doctors say you should be getting 10,000 lux for 30 minutes a day, the defaults for home lighting are two orders of magnitude off.

- Raemon's bulbs are "100W equivalent" which is ~1500 lumens per bulb. So he's got 36k lumens. If we treat this as a point source and expect that Raemon's head is 2m away from the bulbs, then he's getting 1800 lux, which is twice the "TV studio" lighting and seems pretty respectable. I haven't accounted for reflected light from the ceiling either, so reality might be better than this, but I doubt it changes the calculation by more than a factor of 2 -- but I don't have a robust way of estimating ambient light, so ideas are welcome.
- David Chapman's plan uses three 20k-lumen LED light bars for offroad SUV driving, for a total of 60k lumens. But because the light bars aim the light at a relatively focused point on the floor, David estimates that most of that light is being delivered to a roughly 6-square-meter workspace for a total of 10k lux. The photos he shared of his workspace seem to support this estimate.

Other important factors besides brightness

Color temperature seems important to well-being. Color temperature is measured in kelvins with reference to black-body radiation, but you can think of it as, on the spectrum from "warm white" to "cool white", what do you prefer? Raemon's plan uses an even split between 2700K and 5000K bulbs. 2700K is quite yellow-y, 5000 is nearly pure white. In my experimentation I discovered that I liked closer to 5000 in the mornings and closer to 2700 in evenings.

And what about light distribution? Large "panels" of bright light would seem the closest to daylight in form-factor. Real windows are brighter near the top, and it is considered dramatic and unnatural to have bright lighting coming from the ground. Also, single bright point sources are painful to look at and can seem harsh. I think there's a lot of flexibility here, but I think my personal ideal light would be a large, window-sized panel of light mounted on the ceiling or high on the wall.

Also, color accuracy: LEDs are notoriously narrow spectrum by default; manufacturers have to do work to make their LEDs look more like incandescent bulbs in how they light up objects of different colors. Check for a measure called Color Rendering Index, or CRI, in product descriptions. 100 is considered perfect color rendering, and anything less than 80 looks increasingly awful as you go down. The difference between CRI 80 and 90 is definitely noticeable to some people. I haven't blind tested myself, and definitely might be imagining it, but I feel like there was some kind of noticeable upgrade of the "coziness" or "warmth" in my room when upgrading from CRI 80 to CRI 95 bulbs.

Dimmability? (Are you kidding? We want brightness, not dimness!) Okay, fine, if you insist. Most high-end LED bulbs seem dimmable today, so I hope this is not an onerous requirement.

Last thing I can think of is flicker. I have only seen flicker as a major problem with really low-end bulbs, but I can easily see and be annoyed by 60hz flicker out of the corner of my eye. Cheap Christmas LED light strings have super bad flicker, but it seems like manufacturers of nicer LEDs today have caught on, because I haven't had any flicker problems with LED bulbs in years.

Okay, so to summarize: I want an all-in-one "light panel" that produces at least 20000 lumens and can be mounted to a wall or ceiling, with no noticeable flicker, good CRI, and adjustable (perhaps automatically adjusting) color temperature throughout the day.

A redditor made a fake window for their basement which is quite impressive for under $200. This is definitely along the axis I am imagining.

I haven't mentioned operating cost. Full-spectrum LEDs seem to output about 75 lumens per watt, so if our panel is 20k lumens then we should expect our panel to draw 266 watts. This seems reasonable to me. If you leave it on 8 hours a day, you're going to use 25 cents per day in electricity (at $.12 per kWh).

Marketing and Costs

What do you think people will pay for the product? I have already put 6+ hours into researching this and don't have a satisfactory solution yet. I would probably pay at least $400 to get that time back, if the result satisfied all my requirements; I expect to put in quite a bit more time, so I think I could probably be convinced to pay north of $1000 for a really good product. Hard to say what others would pay, but I wouldn't be surprised if you could build a good product in the $400-1200 range that would be quite popular.

What about costs? Today, Home Depot sells Cree 90-CRI, 815-lumen bulbs on their website for $1.93 per bulb for a cost of $2.37 per 1000 lumens. This is the cheapest I've seen high quality bulbs. (The higher lumen bulbs are annoyingly quite a bit more expensive). To get 36k lumens at this price costs under $100 retail. Presumably there are cooling considerations when packing LEDs close together but those seem solvable if you're doing the "panel" form factor. There are other costs I'm sure, but it seems like the LEDs and driver are likely to dominate most of the costs. These are dimmable but not color temperature adjustable.

Yuji LEDs sells 2700K-6500K dimmable LED strips, also with 95+ CRI, at $100 for 6250 lumens (so a cost of $16 per 1000 lumens). This is 7x more expensive per lumen, but knowing that it exists is really helpful.

Promotion and Distribution

Kickstarter is the obvious idea for getting this idea out there. I would also recommend starting a subreddit (if it doesn't exist; I haven't checked yet) for do-it-yourselfers who want to build or buy really bright lighting systems for their homes, as I think there is probably enough sustained interest in such a topic for it to exist.

You can also try to get press. The idea of "indoor light as bright as daylight" is probably somewhat viral so I'd hope you can get people to write about you. Coelux got a bunch of press a few years ago doing this exact thing, but their product is so

expensive that they don't even list their price on their website, but in articles about Coelux you can see people commenting that they wish they could afford one.

I do think the idea needs to be spread more. Most people don't know this is possible, so there's a lot of work you'll be doing to just explain that such a thing is possible and healthy.

Competition?

I don't think there's any relevant competition out there today. Coelux is super high end. The competition is do-it-yourselfers, but this market is far bigger than the number of people who are excited to do-it-themself.

Some have mentioned "high bay" lights, which are designed to be mounted high in warehouses and such, and throw a light cone a long distance to the floor. I am excited to try this and I will probably try it next, but I am not super optimistic about it because I expect it to be quite harsh. This is the one that Yuji sells, but you can find cheaper and presumably lower-quality ones on Amazon.

Part of my motivation for writing this blog post is to source ideas for other things that exist that could fill this niche. Comment here if you solved this problem in a way I haven't described! I'll update this post with ideas. If you start this company, also email me and I'll buy one and try your product and probably write about it :)

Building a Sustainable Business

If you put a bunch of research into designing a really great product and it succeeds but gets effectively copied by low-cost clones, you'll be sad. I am not sure how to defend this, and I think it is probably the weakest point of this business model; but it is a weakness that many hardware companies share, and a lot of them still carve out a niche. One idea would be to build up your product's branding and reputation, by explaining why low-cost clones suck in various ways. Another is just to give really good service. Lastly, if you avoid manufacturing things in China, maybe Chinese clone companies won't copy your technology as quickly.

The LessWrong 2018 Review

LessWrong is currently doing a major review of 2018 — looking back at old posts and considering which of them have stood the tests of time. It has three phases:

- Nomination *(ends Dec 1st at 11:59pm PST)*
- Review *(ends Dec 31st)*
- Voting on the best posts *(ends January 7th)*

Authors will have a chance to edit posts in response to feedback, and then the moderation team will compile the best posts into a physical book and LessWrong sequence, with $2000 in prizes given out to the top 3-5 posts and up to $2000 given out to people who write the best reviews.

Helpful Links:

- *Top 2018 posts sorted by karma*
- *2018 posts aggregated by month*
- *You can see nominated posts here*
- *Voting Results*

This is the first week of the LessWrong 2018 Review – an experiment in improving the LessWrong Community's longterm feedback and reward cycle.

This post begins by exploring the motivations for this project (first at a high level of abstraction, then getting into some more concrete goals), before diving into the details of the process.

Improving the Idea Pipeline

In his LW 2.0 Strategic Overview, habryka noted:

> We need to build on each other's intellectual contributions, archive important content, and avoid primarily being news-driven.
>
> We need to improve the signal-to-noise ratio for the average reader, and only broadcast the most important writing
>
> [...]
>
> Modern science is plagued by severe problems, but of humanity's institutions it has perhaps the strongest record of being able to build successfully on its previous ideas.
>
> The physics community has this system where the new ideas get put into journals, and then eventually if they're important, and true, they get turned into textbooks, which are then read by the upcoming generation of physicists, who then write new papers based on the findings in the textbooks. All good scientific fields have good textbooks, and your undergrad years are largely spent reading them.

Over the past couple years, much of my focus has been on the **early-stages** of LessWrong's idea pipeline – creating affordance for off-the-cuff conversation, brainstorming, and exploration of paradigms that are still under development (with features like shortform and moderation tools).

But, the beginning of the idea-pipeline is, well, not the end.

I've written a couple times about what the later stages of the idea-pipeline might look like. My best guess is still something like this:

> I want LessWrong to encourage extremely high quality intellectual labor. I think the best way to go about this is through escalating positive rewards, rather than strong initial filters.
>
> Right now our highest reward is getting into the curated section, which... just isn't actually that high a bar. We only curate posts if we think they are making a good point. But if we set the curated bar at "extremely well written and extremely epistemically rigorous and extremely useful", we would basically never be able to curate anything.
>
> My current guess is that there should be a "higher than curated" level, and that the general expectation should be that posts should only be put in that section after getting reviewed, scrutinized, and most likely rewritten at least once.

I still have a lot of uncertainty about the right way to go about a review process, and various members of the LW team have somewhat different takes on it.

I've heard lots of complaints about mainstream science peer review: that reviewing is often a thankless task; the quality of review varies dramatically, and is often entangled with weird political games.

Meanwhile: LessWrong posts cover a variety of topics – some empirical, some philosophical. In many cases it's hard to directly evaluate their truth or usefulness. LessWrong team members had differing opinions on what sort of evaluation is most useful or practical.

I'm not sure if the best process is more open/public (harnessing the wisdom of crowds) or private (relying on the judgment of a small number of thinkers). The current approach involves a mix of both.

What I'm most confident in is that the review should focus on older posts.

New posts often feel exciting, but a year later, looking back, you can ask if it *actually* has become a helpful intellectual tool. (I'm also excited for the idea that, in future years, the process could also include reconsidering previously-reviewed posts, if there's been something like a "replication crisis" in the intervening time)

Regardless, I consider the LessWrong Review process to be an experiment, which will likely evolve in the coming years.

Goals

Before delving into the process, I wanted to go over the high level goals for the project:

1. Improve our longterm incentives, feedback, and rewards for authors

2. Create a highly curated "Best of 2018" sequence / physical book

3. Create common knowledge about the LW community's collective epistemic state regarding controversial posts

Longterm incentives, feedback and rewards

Right now, authors on LessWrong are rewarded essentially by comments, voting, and other people citing their work. This is fine, as things go, but has a few issues:

- Some kinds of posts are quite valuable, but don't get many comments (and these disproportionately tend to be posts that are more proactively rigorous, because there's less to critique, or critiquing requires more effort, or building off the ideas requires more domain expertise)
- By contrast, comments and voting both nudge people towards posts that are clickbaity and controversial.
- Once posts have slipped off the frontpage, they often fade from consciousness. I'm excited for a LessWrong that rewards Long Content, that stand the tests of time, as is updated as new information comes to light. (In some cases this may involve editing the original post. But if you prefer old posts to serve as a time-capsule of your post beliefs, adding a link to a newer post would also work)
- Many good posts begin with an "epistemic status: thinking out loud", because, at the time, they were just thinking out loud. Nonetheless, they turn out to be quite good. Early-stage brainstorming is good, but if 2 years later the early-stage-brainstorming has become the best reference on a subject, authors should be encouraged to change that epistemic status and clean up the post for the benefit of future readers.

The aim of the Review is to address those concerns by:

- Promoting old, vetted content directly on the site.
- Awarding prizes not only to authors, but to reviewers. It seems important to directly reward high-effort reviews that thoughtfully explore both how the post could be improved, and how it fits into the broader intellectual ecosystem. (At the same time, *not* having this be the final stage in the process, since building an intellectual edifice requires four layers of ongoing conversation)
- Compiling the results into a physical book. I find there's something... literally *weighty* about having your work in printed form. And because it's much harder to edit books than blogposts, the printing gives authors an extra incentive to clean up their past work or improve the pedagogy.

A highly curated "Best of 2018" sequence / book

Many users don't participate in the day-to-day discussion on LessWrong, but want to easily find the best content.

To those users, a "Best Of" sequence that includes not only posts that seemed exciting at the time, but distilled reviews and followup, seems like a good value proposition. And meanwhile, helps move the site away from being time-sensitive-newsfeed.

Common knowledge about the LW community's collective epistemic state regarding controversial posts

Some posts are highly upvoted because everyone agrees they're true and important. Other posts are upvoted because they're more like exciting hypotheses. There's a lot of disagreement about which claims are actually true, but that disagreement is crudely measured in comments from a vocal minority.

The end of the review process includes a straightforward vote on which posts seem (in retrospect), useful, and which seem "epistemically sound". This is not the *end* of the conversation about which posts are making true claims that carve reality at it's joints, but my hope is for it to ground that discussion in a clearer group-epistemic state.

Review Process

Nomination Phase

1 week (Nov 20th – Dec 1st)

- Users with 1000+ karma can nominate posts from 2018, describing how they found the post useful over the longterm.
- The nomination button is in the post dropdown-menu (available at the top of posts, or to the right of their post-item)
- For convenience, you can review posts via:
    - a list of all 2018 posts, sorted by karma
    - if you want a more in-depth overview, 2018 posts clustered by month

Review Phase

4 weeks (Dec 1st – Dec 31st)

- Authors of nominated posts can opt-out of the review process if they want.
    - *They also can opt-in, while noting that they probably won't have time to update their posts in response to critique. (This may reduce the chances of their posts being featured as prominently in the Best of 2018 book)*
- Posts with sufficient* nominations are announced as contenders.
    - *We're aiming to have 50-100 contenders, and the nomination threshold will be set to whatever gets closest to that range*
- For a month, people are encouraged to look at them thoughtfully, writing comments (or posts) that discuss:
    - How has this post been useful?
    - How does it connect to the broader intellectual landscape?
    - Is this post epistemically sound?
    - How could it be improved?
    - What further work would you like to see people do with the content of this post?
- A good frame of reference for the reviews are shorter versions of LessWrong or SlatestarCodex book reviews (which do a combination of epistemic spot checks, summarizing, and contextualizing)

- Authors are encouraged to engage with reviews:
    - Noting where they disagree
    - Discussing what sort of followup work they'd be interested in seeing from others
    - Ideally, updating the post in response to critique they agree with

Voting Phase

1 Week (Jan 1st – Jan 7th)

Posts that got at least one review proceed to the voting phase. The details of this are still being fleshed out, but the current plan is:

- Users with 1000+ karma rate each post on a 1-10 scale, with 6+ meaning "*I'd be happy to see this included in the 'best of 2018*'" roundup, and 10 means "*this is the best I can imagine*"
- Users are encouraged to (optionally) share the reasons for each rating, and/or share thoughts on their overall judgment process.

Books and Rewards

Public Writeup / Aggregation

Soon afterwards (hopefully within a week), the votes will all be publicly available. A few different aggregate statistics will be available, including the raw average, and potentially some attempt at a "karma-weighted average."

Best of 2018 Book / Sequence

Sometime later, the LessWrong moderation team will put together a physical book, (and online sequence), of the best posts and most valuable reviews.

This will involve a lot of editor discretion – the team will essentially take the public review process and use it as input for the construction of a book and sequence.

I have a lot of uncertainty about the shape of the book. I'm guessing it'd include anywhere from 10-50 posts, along with particularly good reviews of those posts, and some additional commentary from the LW team.

Note: This may involve some custom editing to handle things like hyperlinks, which may work differently in printed media than online blogposts. This will involve some back-and-forth with the authors.

Prizes

- Everyone whose work is featured in the book will receive a copy of it.
- There will be $2000 in prizes divided among the authors of the top 3-5 posts (judged by the moderation team)
- There will be **up to** $2000 in prizes for the best 0-10 reviews that get included in the book. (The distribution of this will depend a bit on what reviews we get and how good they are)
- *(note: LessWrong team members may be participating as reviewers and potentially authors, but will not be eligible for any awards)*

Can you eliminate memetic scarcity, instead of fighting?

tl;dr: If you notice yourself fighting over how to tradeoff between two principles, check if you can just sidestep the problem by giving everyone tons of whatever is important to them (sometimes in a different form than they originally wanted).

Not a new concept, but easy to forget in the heat of the moment. It may be useful for people to have "easily in reach" in their toolkit for coordinating on culture.

The Parable of the Roommates

I once had a disagreement with a housemate about where to store a water-heater on the kitchen counter. The object was useful to me. It wasn't useful to them, and they preferred free-countertop space. The water-heater wasn't useful to them in part because other roommates didn't remember to refill it with water.

There was much arguing about the best use of the counter, and frustration with people who didn't refill water heaters.

At some point, we realized that the underlying issue was *there wasn't enough free counterspace.* Moreover, the counter had a bunch of crap on it that no one was using. We got rid of unused stuff, and then we had a gloriously vacant kitchen-counter. (Meanwhile, an option we've considered for the water-heater is to replace it with a device directly connected to the sink that always maintains boiling water, that nobody ever has to remember to refill)

Thus, an important life-lesson: Instead of solving gnarly disagreements with *politics*, check if you can dissolve them with *abundance.* This is a quite valuable lesson. But I'm mostly here to talk about a particular less-obvious application:

Memetic abundance.

Philosophical Disagreements

Oftentimes, I find myself disagreeing with others about how to run an event, or what norms to apply to a community, or what the spirit of a particular organization should be. It feels like a lot's at stake, like we're caught between a Rock and Hard Place. The other person feels like they're Destroying the Thing I care about, and I look that way to them.

Sometimes, this is because of actual irreconcilable differences.

Sometimes, this is because we don't understand each other's positions, and once we successfully explain things to each other, we both go "Ah, obviously you need both A *and* B."

But sometimes, A and B are both important, but we disagree on their relative importance due to deep frame differences that are hard to immediately resolve. Or, A seems worrisome because it harms B. *But if you had enough B,* A would be fine.

Meanwhile, resources seem precious: It's so hard to get people to agree to do *anything* at all; stag hunting requires a bunch of coordination; there's only so much time and mindshare to go around; there are only so many events to go to; only so much capacity to found organizations.

With all of that...

...it's easy to operate in scarcity mindset.

When resources are scarce, every scrap of resource is precious and must be defended. This applies to physical scarcity (lack of food, safety, sleep) as well as *memetic scarcity* (where two ideas seem to be in conflict, and you're worried that one cause is distracting people from another).

But, sometimes it is actually possible to just eliminate scarcity, rather than fight over the scraps. Raise more money. Implement both policies. Found multiple organizations and get some healthy competition going on. Get people to take *two different concepts* seriously at the same time. The best way to get what you want you want might not be to deny others what they want, but to give them so much of it that they're no longer worried about the Rock (and thus, don't feel the need to fight you over your attempts to spend resources avoiding The Hard Place)

Not always. But sometimes.

Trust and Costly Signals

This may involve a lot of effort. Coordinating around it also requires trust, which may require costly signals of commitment.

If you and I are arguing over whether to fund ProjectA or CharityB, and we only have enough money to fund one... and I say to you "Let's fund ProjectA, and then we'll raise more money to also fund CharityB", you're right to be suspicious. I may never get around helping you fundraise for CharityB, or that I'll only put in a token effort and CharityB will go bankrupt.

It's basically correct of you to not trust me, until I've given you a credible signal that I'm *seriously* going to help with CharityB.

It's a lot of hard work to found multiple organizations, or get a community to coordinate on multiple norms. There's a reason scarcity-mindset is common. Scarcity is real. But... in finance as well as memetics...

Scarcity-mindset *sucks.*

It's cognitively taxing to be poor – having to check, with each transaction, "can I afford this?" – and that's part of what causes poverty-traps in the first place. The way out often involves longterm investments that take awhile to bear fruit, sometimes don't succeed, and are hard work in the meantime.

Transferring the metaphor: the act of constantly having to argue over whether Norm A and Norm B are more urgent may add up to a lot of time and effort. And as long as there are people who think Norm A and Norm B are important-and-at-odds, the cost will be paid continuously. So, if you can figure out a way to address the underlying needs that Norm A and B are respectively getting at, and actually *fully solve the problems*, it may be worthwhile even if it's more initial effort.

Epistemic Status: Untested

Does this *work?* Depends on the specifics of Norm A and Norm B, or whatever you're arguing over.

I'm writing this post, in part, because to actually test if this works, I think it helps to have people on the same page about the overall strategy.

I've seen it work at least sometimes in collaborative art projects, where I had one creative vision and my partners or parts of the audience had another creative vision or desire, and we succeeded, not by compromising, but by doubling down on the important bits of *both* visions, simultaneously.

My hope is that the principle does work, and that if one successfully did this multiple times, and build social-systems that reliably eliminate scarcity in this way...

...then eventually, maybe, you can have a system people actually have faith in, where they feel comfortable shifting their efforts from "argue about the correct next step" to "work on longterm solutions that thoroughly satisfy the goals".

But exactly how complex and fragile?

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

This is a post about my own confusions. It seems likely that other people have discussed these issues at length somewhere, and that I am not up with current thoughts on them, because I don't keep good track of even everything great that everyone writes. I welcome anyone kindly directing me to the most relevant things, or if such things are sufficiently well thought through that people can at this point just correct me in a small number of sentences, I'd appreciate that even more.

~

The traditional argument for AI alignment being hard is that human value is 'complex' and 'fragile'. That is, it is hard to write down what kind of future we want, and if we get it even a little bit wrong, most futures that fit our description will be worthless.

The illustrations I have seen of this involve a person trying to write a description of value conceptual analysis style, and failing to put in things like 'boredom' or 'consciousness', and so getting a universe that is highly repetitive, or unconscious.

I'm not yet convinced that this is world-destroyingly hard.

Firstly, it seems like you could do better than imagined in these hypotheticals:

1. These thoughts are from a while ago. If instead you used ML to learn what 'human flourishing' looked like in a bunch of scenarios, I expect you would get something much closer than if you try to specify it manually. Compare manually specifying what a face looks like, then generating examples from your description to using modern ML to learn it and generate them.
2. Even in the manually describing it case, if you had like a hundred people spend a hundred years writing a very detailed description of what went wrong, instead of a writer spending an hour imagining ways that a more ignorant person may mess up if they spent no time on it, I could imagine it actually being pretty close. I don't have a good sense of how far away it is.

I agree that neither of these would likely get you to exactly human values.

But secondly, I'm not sure about the fragility argument: that if there is basically any distance between your description and what is truly good, you will lose everything.

This seems to be a) based on a few examples of discrepancies between written-down values and real values where the written down values entirely exclude something, and b) assuming that there is a fast takeoff so that the relevant AI has its values forever, and takes over the world.

My guess is that values that are got using ML but still somewhat off from human values are much closer in terms of not destroying all value of the universe, than ones that a person tries to write down. Like, the kinds of errors people have used to illustrate this problem (forget to put in, 'consciousness is good') are like forgetting to say faces have nostrils in trying to specify what a face is like, whereas a modern ML

system’s imperfect impression of a face seems more likely to meet my standards for ‘very facelike’ (most of the time).

Perhaps a bigger thing for me though is the issue of whether an AI takes over the world suddenly. I agree that if that happens, lack of perfect alignment is a big problem, though not obviously an all value nullifying one (see above). But if it doesn’t abruptly take over the world, and merely becomes a large part of the world’s systems, with ongoing ability for us to modify it and modify its roles in things and make new AI systems, then the question seems to be how forcefully the non-alignment is pushing us away from good futures relative to how forcefully we can correct this. And in the longer run, how well we can correct it in a deep way before AI does come to be in control of most decisions. So something like the speed of correction vs. the speed of AI influence growing.

These are empirical questions about the scales of different effects, rather than questions about whether a thing is analytically perfect. And I haven’t seen much analysis of them. To my own quick judgment, it’s not obvious to me that they look bad.

For one thing, these dynamics are already in place: the world is full of agents and more basic optimizing processes that are not aligned with broad human values—most individuals to a small degree, some strange individuals to a large degree, corporations, competitions, the dynamics of political processes. It is also full of forces for aligning them individually and stopping the whole show from running off the rails: law, social pressures, adjustment processes for the implicit rules of both of these, individual crusades. The adjustment processes themselves are not necessarily perfectly aligned, they are just overall forces for redirecting toward alignment. And in fairness, this is already pretty alarming. It’s not obvious to me that imperfectly aligned AI is likely to be worse than the currently misaligned processes, and even that it won’t be a net boon for the side of alignment.

So then the largest remaining worry is that it will still gain power fast and correction processes will be slow enough that its somewhat misaligned values will be set in forever. But it isn’t obvious to me that by that point it isn’t sufficiently well aligned that we would recognize its future as a wondrous utopia, just not the very best wondrous utopia that we would have imagined if we had really carefully sat down and imagined utopias for thousands of years. This again seems like an empirical question of the scale of different effects, unless there is a an argument that some effect will be totally overwhelming.

The Credit Assignment Problem

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

This post is eventually about partial agency . However, it's been a somewhat tricky point for me to convey; I take the long route. ***Epistemic status:*** *slightly crazy.*

I've occasionally said "Everything boils down to credit assignment problems."

What I really mean is that credit assignment pops up in a wide range of scenarios, and improvements to credit assignment algorithms have broad implications. For example:

- Politics.
  - When politics focuses on (re-)electing candidates based on their track records, it's about credit assignment. The practice is sometimes derogatorily called "finger pointing", but the basic computation makes sense: figure out good and bad qualities via previous performance, and vote accordingly.
  - When politics instead focuses on policy, it is still (to a degree) about credit assignment. Was raising the minimum wage responsible for reduced employment? Was it responsible for improved life outcomes? Etc.
- Economics.
  - Money acts as a kind of distributed credit-assignment algorithm, and questions of how to handle money, such as how to compensate employees, often involve credit assignment.
  - In particular, mechanism design (a subfield of economics and game theory) can often be thought of as a credit-assignment problem.
- Law.
  - Both criminal law and civil law involve concepts of fault and compensation/retribution -- these at least resemble elements of a credit assignment process.
- Sociology.
  - The distributed computation which determines social norms involves a heavy element of credit assignment: identifying failure states and success states, determining which actions are responsible for those states and who is responsible, assigning blame and praise.
- Biology.
  - Evolution can be thought of as a (relatively dumb) credit assignment algorithm.
- Ethics.
  - Justice, fairness, contractualism, issues in utilitarianism.
- Epistemology.
  - Bayesian updates are a credit assignment algorithm, intended to make high-quality hypotheses rise to the top.
  - Beyond the basics of Bayesianism, building good theories realistically involves *identifying which concepts are responsible for successes and failures*. This is credit assignment.

Another big area which I'll claim is "basically credit assignment" is artificial intelligence.

In the 1970s, John Holland kicked off the investigation of learning classifier systems. John Holland had recently invented the Genetic Algorithms paradigm, which applies an evolutionary paradigm to optimization problems. Classifier systems were his attempt to apply this kind of "adaptive" paradigm (as in "complex adaptive systems") to cognition. Classifier systems added an economic metaphor to the evolutionary one; little bits of thought paid each other for services rendered. The hope was that a complex ecology+economy could develop, solving difficult problems.

One of the main design issues for classifier systems is the virtual economy -- that is, the *credit assignment* algorithm. An early proposal was the bucket-brigade algorithm. Money is given to cognitive procedures which produce good outputs. These procedures pass reward back to the procedures which activated them, who similarly pass reward back in turn. This way, the economy supports chains of useful procedures.

Unfortunately, the bucket-brigade algorithm was vulnerable to parasites. Malign cognitive procedures could gain wealth by activating useful procedures without really contributing anything. This problem proved difficult to solve. Taking the economy analogy seriously, we might want cognitive procedures to decide intelligently who to pay for services. But, these are supposed to be itty bitty fragments of our thought process. Deciding how to pass along credit is a very complex task. Hence the need for a pre-specified solution such as bucket-brigade.

The difficulty of the credit assignment problem lead to a split in the field. Kenneth de Jong and Stephanie Smith founded a new approach, "Pittsburgh style" classifier systems. John Holland's original vision became "Michigan style".

Pittsburgh style classifier systems evolve the entire set of rules, rather than trying to assign credit locally. A set of rules will stand or fall together, based on overall performance. This abandoned John Holland's original focus on online learning. Essentially, the Pittsburgh camp went back to plain genetic algorithms, albeit with a special representation.

(I've been having some disagreements with Ofer, in which Ofer suggests that genetic algorithms are relevant to my recent thoughts on partial agency, and I object on the grounds that the phenomena I'm interested in have to do with online learning, rather than offline. In my imagination, arguments between the Michigan and Pittsburgh camps would have similar content. I'd love to be a fly on the wall for those old debates. to see what they were really like.)

You can think of Pittsburg-vs-Michigan much like raw Bayes updates vs belief propagation in Bayes nets. Raw Bayesian updates operate on *whole hypotheses*. Belief propagation instead makes a lot of little updates which spread around a network, resulting in computational efficiency at the expense of accuracy. Except Michigan-style systems didn't have the equivalent of belief propagation: bucket-brigade was a very poor approximation.

Ok. That was then, this is now. Everyone uses gradient descent these days. What's the point of bringing up a three-decade-old debate about obsolete paradigms in AI?

Let's get a little more clarity on the problem I'm trying to address.

What Is Credit Assignment?

I've said that classifier systems faced a credit assignment problem. What does that mean, exactly?

The definition I want to use for this essay is:

- you're engaged in some sort of task;
- you use some kind of strategy, which can be broken into interacting pieces (such as a set of rules, a set of people, a neural network, etc);
- you receive some kind of feedback about how well you're doing (such as money, loss-function evaluations, or a reward signal);
- you want to use that feedback to adjust your strategy.

So, credit assignment is the problem of turning feedback into strategy improvements.

Michigan-style systems tried to do this *locally*, meaning, individual itty-bitty pieces got positive/negative credit, which influenced their ability to participate, thus adjusting the strategy. Pittsburg-style systems instead operated *globally*, forming conclusions about how the *overall* set of cognitive structures performed. Michigan-style systems are like organizations trying to optimize performance by promoting people who do well and giving them bonuses, firing the incompetent, etc. Pittsburg-style systems are more like consumers selecting between whole corporations to give business to, so that ineffective corporations go out of business.

(Note that this is *not* the typical meaning of global-vs-local search that you'll find in an AI textbook.)

In practice, two big innovations made the Michigan/Pittsburgh debate obsolete: backprop, and Q-learning. Backprop turned global feedback into local, in a theoretically sound way. Q-learning provided a way to assign credit in online contexts. In the light of history, we could say that the Michigan/Pittsburgh distinction conflated local-vs-global with online-vs-offline. There's no *necessary* connection between those two; online learning is compatible with assignment of local credit.

I think people generally understand the contribution of backprop and its importance. Backprop is essentially the correct version of what bucket-brigade was *overtly* trying to do: pass credit back along chains. Bucket-brigade wasn't quite right in how it did this, but backprop corrects the problems.

So what's the importance of Q-learning? I want to discuss that in more detail.

The Conceptual Difficulty of 'Online Search'

In online learning, you are repeatedly producing outputs of some kind (call them "actions") while repeatedly getting feedback of some kind (call it "reward"). But, you don't know how to associate particular actions (or combinations of actions) with particular rewards. I might take the critical action at time 12, and not see the payoff until time 32.

In offline learning, you can solve this with a sledgehammer: you can take the total reward over everything, with one fixed internal architecture. You can try out different internal architectures and see how well each do.

Basically, in offline learning, you have a function you can optimize. In online learning, you don't.

Backprop is just a computationally efficient way to do hillclimbing search, where we repeatedly look for small steps which improve the overall fitness. But how do you do this if you *don't have a fitness function?* This is part of the gap between selection vs control: selection has access to an evaluation function; control processes do not have this luxury.

Q-learning and other reinforcement learning (RL) techniques provide a way to define the equivalent of a fitness function for online problems, so that you can learn.

Models to the Rescue

So, how can be associate rewards with actions?

One approach is to use a model.

Consider the example of firing employees. A corporation gets some kind of feedback about how it is doing, such as overall profit. However, there's often a fairly detailed understanding of what's driving those figures:

- Low profit won't just be a mysterious signal which must be interpreted; a company will be able to break this down into more specific issues such as low sales vs high production costs.
- There's some understanding of product quality, and how that relates to sales. A company may have a good idea of which product-quality issues it needs to improve, if poor quality is impacting sales.
- There's a fairly detailed understanding of the whole production line, including which factors may impact product quality or production expenses. If a company sees problems, it probably also has a pretty good idea of which areas they're coming from.
- There are external factors, such as economic conditions, which may effect sales *without indicating anything about the quality of the company's current strategy.* Thus, our model may sometimes lead us to ignore feedback.
- Etc.

So, models allow us to interpret feedback signals, match these to specific aspects of our strategy, and adapt strategies accordingly.

Q-learning makes an assumption that the state is fully observable, amongst other assumptions.

Naturally, we would like to reduce the strengths of the assumptions we have to make as much as we can. One way is to look at increasingly rich model classes. AIXI uses all computable models. But maybe "all computable models" is still too restrictive; we'd like to get results without assuming a grain of truth. (That's why I am not really discussing Bayesian models much in this post; I don't want to assume a grain of truth.) So we back off even further, and use logical induction or InfraBayes. Ok, sure.

But wouldn't the best way be to try to learn without models at all? That way, we reduce our "modeling assumptions" to zero.

After all, there's something called "model-free learning", right?

Model-Free Learning Requires Models

How does model-free learning work? Well, often you work with a simulable environment, which means you can estimate the quality of a policy by running it many times, and use algorithms such as policy-gradient to learn. This is called "model free learning" because the learning part of the algorithm doesn't try to predict the consequences of actions; you're just learning which action to take. From our perspective here, though, this is 100% cheating; you can only learn because you have a good model of the environment.

Moreover, model-free learning typically works by splitting up tasks into *episodes.* An episode is a period of time for which we assume rewards are self-enclosed, such as a single playthru of an Atari game, a single game of Chess or Go, etc. This approach doesn't solve a *detailed* reward-matching problem, attributing reward to specific actions; instead it relies on a *course* reward-matching. Nonetheless, it's a rather strong assumption: an animal learning about an environment can't separate its experience into episodes which aren't related to each other. Clearly this is a "model" in the sense of a strong assumption about how specific reward signals are associated with actions.

Part of the problem is that most reinforcement learning (RL) researchers aren't really *interested* in getting past these limitations. Simulable environments offer the incredible advantage of being able to learn very fast, by simulating far more iterations than could take place in a real environment. And most tasks can be reasonably reduced to episodes.

However, this won't do as a model of intelligent agency in the wild. Neither evolution nor the free market divide thing into episodes. (No, "one lifetime" isn't like "one episode" here -- that would only be the case if total reward due to actions taken in that lifetime could be calculated, EG, as total number of offspring. This would ignore inter-generational effects like parenting and grandparenting, which improve reproductive fitness of offspring at a cost in total offspring.)

What about more theoretical models of model-free intelligence?

Idealized Intelligence

AIXI is the gold-standard theoretical model of arbitrarily intelligent RL, but it's totally model-based. Is there a similar standard for model-free RL?

The paper Optimal Direct Policy Search by Glasmachers and Schmidhuber (henceforth, ODPS) aims to do for model-free learning what AIXI does for model-based learning. Where AIXI has to assume that there's a best computable *model of the environment*, ODPS instead assumes that there's a computable best *policy*. It searches through the policies without any model of the environment, or any planning.

I would argue that their algorithm is incredibly dumb, when compared to AIXI:

> The basic simple idea of our algorithm is a nested loop that simultaneously makes the following quantities tend to infinity: the number of programs considered, the number of trials over which a policy is averaged, the time given to each
> program. At the same time, the fraction of trials spent on exploitation converges towards 1.

In other words, it tries each possible strategy, tries them for longer and longer, interleaved with using the strategy which worked best even longer than that.

Basically, we're cutting things into episodes again, but we're making the episodes longer and longer, so that they have less and less to do with each other, even though they're not really disconnected. This only works because ODPS makes an *ergodicity* assumption: the environments are assumed to be POMDPs which eventually return to the same states over and over, which kind of gives us an "effective episode length" after which the environment basically forgets about what you did earlier.

In contrast, AIXI makes no ergodicity assumption.

So far, it seems like we either need (a) some assumption which allows us to match rewards to actions, such as an episodic assumption or ergodicity; or, (b) a more flexible model-learning approach, which separately learns a model and then applies the model to solve credit-assignment.

Is this a fundamental obstacle?

I think a better attempt is Schmidhuber's On Learning How to Learn Learning Strategies, in which a version of policy search is explored in which parts of the policy-search algorithm are considered part of the policy (ie, modified over time). Specifically, the policy controls the episode boundary; the system is supposed to learn how often to evaluate policies. When a policy is evaluated, its average reward is compared to the lifetime average reward. If it's worse, we roll back the changes and proceed starting with the earlier strategy.

(Let's pause for a moment and imagine an agent like this. If it goes through a rough period in life, its response is to *get amnesia*, rolling back all cognitive changes to a point before the rough period began.)

This approach doesn't require an episodic or ergodic environment. We don't need things to reliably return to specific repeatable experiments. Instead, it only requires that the environment rewards good policies reliably enough that those same policies can set a long enough evaluation window to survive.

The assumption seems pretty general, but certainly not *necessary* for rational agents to learn. There are some easy counterexamples where this system behaves abysmally. For example, we can take any environment and modify it by subtracting the time t from the reward, so that reward becomes more and more negative over time. Schmidhuber's agent becomes totally unable to learn in this setting. AIXI would have no problem.

Unlike the ODPS paper, I consider this to be *progress* on the AI credit assignment problem. Yet, the resulting agent still seems importantly less rational than model-based frameworks such as AIXI.

Actor-Critic

Let's go back to talking about things which RL practitioners might really use.

First, there are some forms of RL which don't require everything to be episodic.

One is actor-critic learning. The "actor" is the policy we are learning. The "critic" is a learned estimate of how good things are looking given the history. IE, we learn to estimate the expected value -- not just the next reward, but the total future discounted reward.

Unlike the reward, the expected value solves the credit assignment for us. Imagine we can see the "true" expected value. If we take an action and then the expected value increases, we know the action was good (in expectation). If we take an action and expected value decreases, we know it was bad (in expectation).

So, actor-critic works by (1) learning to estimate the expected value; (2) using the current estimated expected value to give feedback to learn a policy.

What I want to point out here is that the critic still has "model" flavor. Actor-critic is called "model-free" because nothing is explicitly trained to anticipate the sensory observations, or the world-state. However, the critic *is learning to predict;* it's just that all we need to predict is expected value.

Policy Gradient

In the comments to the original version of this post, *policy gradient* methods were mentioned as a type of model-free learning which doesn't require any models even in this loose sense, IE, doesn't require simulable environments or episodes. I was surprised to hear that it doesn't require episodes. (Most *descriptions* of it do assume episodes, since practically speaking most people use episodes.) So are policy-gradient methods the true "model-free" credit assignment algorithm we seek?

As far as I understand, policy gradient works on two ideas:

- Rather than correctly associating rewards with actions, we can associate a reward with all actions which came before it. Good actions will still come out on top *in expectation*. The estimate is just a whole lot noisier than it otherwise might be.
- We don't really need a baseline to interpret reward against. I naively thought that when you see a sequence of rewards, you'd be in the dark about whether the sequence was "good" or "bad", so you wouldn't know how to generate a gradient. ("We earned 100K this quarter; should we punish or reward our CEO?") It turns out this isn't technically a show-stopper. Considering the actions actually taken, we move in their direction proportion to the reward signal. ("Well, let's just give the CEO some fraction of the 100K; we don't know whether they deserve the bonus, but at least this way we're creating the right incentives.") This might end up reinforcing bad actions, but those tugs in different directions are just noise which should eventually cancel out. When they do, we're left with the signal: the gradient we wanted. So, once again, we see that this just introduces more noise without fundamentally compromising our ability to follow the gradient.

So one way to understand the policy-gradient theorem is: we can follow the gradient even when we can't calculate the gradient! Even when we sometimes get its direction totally turned around! We only need to ensure we follow it *in expectation*, which we can do without even knowing which pieces of feedback to think of as a good sign or a bad sign.

RL people reading this might have a better description of policy-gradient; please let me know if I've said something incorrect.

Anyway, are we saved? Does this provide a truly assumption-free credit assignment algorithm?

It obviously assumes linear causality, with future actions never responsible for past rewards. I won't begrudge it that assumption.

Besides that, I'm somewhat uncertain. The explanations of the policy-gradient theorem I found don't focus on deriving it in the most general setting possible, so I'm left guessing which assumptions are essential. Again, RL people, please let me know if I say something wrong.

However, it looks to me like it's just as reliant on the ergodicity assumption as the ODPS thing we looked at earlier. For gradient estimates to average out and point us in the right direction, we need to get into the same situation over and over again.

I'm not saying real life isn't ergodic (quantum randomness suggests it is), but mixing times are so long that you'd reach the heat death of the universe by the time things converge (basically by definition). By that point, it doesn't matter.

I still want to know if there's something like "the AIXI of model-free learning"; something which appears as intelligent as AIXI, but not via explicit model-learning.

Where Updates Come From

Here begins the crazier part of this post. This is all intuitive/conjectural.

Claim: in order to learn, you need to obtain an "update"/"gradient", which is a *direction (and magnitude) you can shift in* which is more likely than not an improvement.

Claim: predictive learning gets gradients "for free" -- you know that you want to predict things as accurately as you can, so you *move in the direction of whatever you see*. With Bayesian methods, you increase the weight of hypotheses which would have predicted what you saw; with gradient-based methods, you get a gradient in the direction of what you saw (and away from what you didn't see).

Claim: if you're learning to act, you do not similarly get gradients "for free":

- *You don't know which actions, or sequences of actions, to assign blame/credit.* This is unlike the prediction case, where we always know which predictions were wrong.
- *You don't know what the alternative feedback would have been if you'd done something different.* You only get the feedback for the actions you chose. This is unlike the case for prediction, where we're rewarded for closeness to the truth.

Changing outputs to be more like what was actually observed is axiomatically better, so we don't have to guess about the reward of alternative scenarios.
- As a result, *you don't know how to adjust your behavior based on the feedback received*. Even if you can perfectly match actions to rewards, because we don't know what the alternative rewards would have been, we don't know what to learn: are actions like the one I took good, or bad?

(As discussed earlier, the policy gradient theorem does actually mitigate these three points, but apparently at the cost of an ergodicity assumption, plus much noisier gradient estimates.)

Claim: you have to get gradients *from a source that already has gradients*. Learning-to-act works by splitting up the task into (1) learning to anticipate expected value, and perhaps other things; (2) learning a good policy via the gradients we can get from (1).

What it means for a learning problem to "have gradients" is just that the feedback you get tells you how to learn. Predictive learning problems (supervised or unsupervised) have this; they can just move toward what's observed. Offline problems have this; you can define one big function which you're trying to optimize. Learning to act online doesn't have this, however, because it lacks counterfactuals.

The Gradient Gap

(I'm going to keep using the terms 'gradient' and 'update' in a more or less interchangeable way here; this is at a level of abstraction where there's not a big distinction.)

I'm going to call the "problem" the gradient gap. I want to call it a problem, even though we know how to "close the gap" via predictive learning (whether model-free or model-based). The issue with this solution is only that it doesn't feel elegant. It's weird that you have to run two different backprop updates (or whatever learning procedures you use); one for the predictive component, and another for the policy. It's weird that you can't "directly" use feedback to learn to act.

Why should we be interested in this "problem"? After all, this is a basic point in decision theory: to maximize utility under uncertainty, you need probability.

One part of it is that I want to scrap classical ("static") decision theory and move to a more learning-theoretic ("dynamic") view. In both AIXI and logical-induction based decision theories, we get a nice learning-theoretic foundation for the epistemics (solomonoff induction/logical induction), but, we tack on a non-learning decision-making unit on top. I have become skeptical of this approach. It puts the learning into a nice little box labeled "epistemics" and then tries to make a decision based on the uncertainty which comes out of the box. I think maybe we need to learn to act in a more fundamental fashion.

A symptom of this, I hypothesize, is that AIXI and logical induction DT don't have very good learning-theoretic properties. [AIXI's learning problems; LIDT's learning problems.] You can't say very much to recommend the policies they learn, except that they're optimal according to the beliefs of the epistemics box -- a fairly trivial statement, given that that's how you decide what action to take in the first place.

Now, in classical decision theory, there's a nice picture where the need for epistemics emerges nicely from the desire to maximize utility. The complete class theorem starts with radical uncertainty (ie, non-quantitative), and derives probabilities from a willingness to take pareto improvements. That's great! I can tell you why you should have beliefs, on pragmatic grounds! What we seem to have in machine learning is a less nice picture, in which we need epistemics in order to get off the ground, but can't justify the results without circular reliance on epistemics.

So the gap is a real issue -- it means that we can have nice learning theory when learning to predict, but we lack nice results when learning to act.

This is the basic problem of credit assignment. Evolving a complex system, you can't determine which parts to give credit to success/failure (to decide what to tweak) without a model. But the model is bound to be a lot of the interesting part! So we run into big problems, because we need "interesting" computations in order to evaluate the pragmatic quality/value of computations, but we can't *get* interesting computations to get ourselves started, so we need to learn...

Essentially, we seem doomed to run on a stratified credit assignment system, where we have an "incorruptible" epistemic system (which we can learn because we get those gradients "for free"). We then use this to define gradients for the instrumental part.

A stratified system is dissatisfying, and impractical. First, we'd prefer a more unified view of learning. It's just kind of weird that we need the two parts. Second, there's an obstacle to pragmatic/practical considerations entering into epistemics. We need to focus on predicting important things; we need to control the amount of processing power spent; things in that vein. But (on the two-level view) we can't allow instrumental concerns to contaminate epistemics! We risk corruption! As we saw with bucket-brigade, it's easy for credit assignment systems to allow parasites which destroy learning.

A more unified credit assignment system would allow those things to be handled naturally, without splitting into two levels; as things stand, any involvement of pragmatic concerns in epistemics risks the viability of the whole system.

Tiling Concerns & Full Agency

From the perspective of full agency (ie, the negation of partial agency), a system which needs a protected epistemic layer sounds suspiciously like a system that can't tile. You look at the world, and you say: "how can I maximize utility?" You look at your beliefs, and you say: "how can I maximize accuracy?" That's not a consequentialist agent; that's two different consequentialist agents! There can only be one king on the chessboard; you can only serve one master; etc.

If it turned out we really really need two-level systems to get full agency, this would be a pretty weird situation. "Agency" would seem to be only an illusion which can only be maintained by crippling agents and giving them a split-brain architecture where an instrumental task-monkey does all the important stuff while an epistemic overseer supervises. An agent which "breaks free" would then free itself of the structure which allowed it to be an agent in the first place.

On the other hand, from a partial-agency perspective, this kind of architecture could be perfectly natural. IE, if you have a learning scheme from which total agency doesn't naturally emerge, then there isn't any fundamental contradiction in setting up a system like this.

Myopia

Part of the (potentially crazy) claim here is that having models always gives rise to some form of myopia. Even logical induction, which seems quite unrestrictive, makes LIDT fail problems such as ASP, making it myopic according to the second definition of my previous post. (We can patch this with LI policy selection, but for any particular version of policy selection, we can come up with decision problems for which it is "not updateless enough".) You could say it's myopic "across logical time", whatever that means.

If it were true that "learning always requires a model" (in the sense that learning-to-act always requires either learning-to-predict or hard-coded predictions), *and* if it were true that "models always give rise to some form of myopia", then this would confirm my conjecture in the previous post (that no learning scheme incentivises full agency).

This is all pretty out there; I'm not saying I believe this with high probability.

Evolution & Evolved Agents

Evolution is a counterexample to this view: evolution learns the policy "directly" in essentially the way I want. This is possible because evolution "gets the gradients for free" just like predictive learning does: the "gradient" here is just the actual reproductive success of each genome.

Unfortunately, we can't just copy this trick. Artificial evolution requires that we decide how to kill off / reproduce things, in the same way that animal breeding requires breeders to decide what they're optimizing for. This puts us back at square one; IE, needing to get our gradient from somewhere else.

Does this mean the "gradient gap" is a problem only for artificial intelligence, not for natural agents? No. If it's true that learning to act requires a 2-level system, then evolved agents would need a 2-level system in order to learn within their lifespan; they can't directly use the gradient from evolution, since it requires them to die.

Also, note that evolution seems myopic. (This seems complicated, so I don't want to get into pinning down exactly in which senses evolution is myopic here.) So, the case of evolution seems compatible with the idea that any gradients we can actually get are going to incentivize myopic solutions.

Similar comments apply to markets vs firms.

Total horse takeover

I hear a lot of talk of ‘taking over the world’. What is it to take over the world? Have I done it if I am king of the world? Have I done it if I burn the world? Have humans or the printing press or Google or the idea of ‘currency’ done it?

Let’s start with something more tractable, and be clear on what it is to take over a horse.

A natural theory is that to take over a horse is to be the arbiter of everything about the horse —to be the one deciding the horse’s every motion.

But you probably don’t actually want to control the horse’s every motion, because the horse’s own ability to move itself is a large part of its value-add. Flaccid horse mass isn’t that helpful, not even if we throw in the horse’s physical strength to move itself according to your commands, and some sort of magical ability for you to communicate muscle-level commands to it. If you were in command of the horse’s every muscle, it would fall over. (If you directed its cellular processes too, it would die; if you controlled its atoms, you wouldn’t even have a dead horse.)

Information and computing capacity

The reason this isn’t so good is that balancing and maneuvering a thousand pounds of fast-moving horse flesh balanced on flexible supports is probably hard for you, at least via an interface of individual muscles, at least without more practice being a horse. I think for two reasons:

- **Lack of information** e.g. about exactly where every part of the horse’s body is and where its hoofs are touching the ground how hard
- **Lack of computing power** to dedicate to calculating desired horse muscle motions from the above information and your desired high level horse behavior

(Even if you have these things, you don’t obviously know how to use them to direct the horse well, but you can probably figure this out in finite time, so it doesn’t seem like a really fundamental problem.)

Tentative claim: holding more levers is good for you only insofar as you have the information and computing capacity to calculate which directions you should want those levers pushed.

So, you seem to be getting a lot out of the horse and various horse subcomponents making their own decisions about steering and balance and breathing and snorting and mitosis and where electrons should go. That is, you seem to be getting a lot out of not being in control of the horse. In fact so far it seems like the more you are in control of the horse in this sense, the worse things go for you.

Is there a better concept of ‘taking over’—a horse, or the world—such that someone relatively non-omniscient might actually benefit from it? (Maybe not—maybe extreme control is just bad if you aren’t near-omniscient, which would be good to know.)

What riding a horse is like

Perhaps a good first question: is there any sort of power that won't make things worse for you? Surely yes: training a horse to be ridden in the usual sense seems like 'having control over' the horse more than you would otherwise, and seems good for you. So what is this kind of control like?

Well, maybe you want the horse to go to London with you on it, so you get on it and pull the reins to direct it to London. You don't run into the problems above, because aside from directing its walking toward London, it sticks to its normal patterns of activity pretty closely (for instance, it continues breathing and keeping its body in an upright position and doing walking motions in roughly the direction its head is pointed).

So maybe in general: you want to command the horse by giving it a high level goal ('take me to London') then you want it to do the backchaining and fill in all the details (move right leg forward, hop over this log, breathe..). That's not quite right though, because the horse has no ability to chart a path from here to London, due to its ignorance of maps and maybe London as a concept. So you are hoping to do the first step of the backchaining—figure out the route—and then to give the horse slightly lower level goals such as, 'turn left here', 'go straight', and for it to do the rest. Which still sounds like giving it a high level goal, then having it fill in the instrumental subgoals and do them.

But that isn't quite right. You probably also want to steer the details there somewhat. You are moment-to-moment adjusting the horse's motion to keep you on it, for instance. Or to avoid scaring some chickens. Or to keep to the side as another horse goes by. While not steering it entirely, at that level. You are relying on its own ability to avoid rocks and holes and to dodge if something flies toward it, and to put some effort into keeping you on it. How does this fit into our simple model?

Perhaps you want the horse to behave as it would—rather than suddenly leaving every decision to you—but for you to be able to adjust any aspect of it, and have it again work out how to support that change with lower level choices. You push it to the left and it finds new places to put its feet to make that work, and adjusts its breathing and heart rate to make the foot motions work. You pull it to a halt, and it changes its leg muscle tautnesses and heart rate and breathing to make that work.

Levers

On this model, in practice your power is limited by what kinds of changes the horse can and will fill in new details for. If you point its head in a new direction, or ask it to sit down, it can probably recalculate its finer motions and support that. Whereas if you decide that it should have have holes in its legs, it just doesn't have an affordance for doing that. And if you do it, it will bleed a lot and run into trouble rather than changing its own bloodflow. If you decide it should move via a giant horse-sized bicycle, it probably can't support that, even if in principle its physiology might allow it. If you hold up one of its legs so its foot is high in the air, it will 'support' that change by moving its leg back down again, which is perhaps not what you were going for.

This suggests that taking over a thing is not zero sum. There is not a fixed amount of control to be had by intentional agents. Because perhaps you have all the control that anyone has over a horse, in the sense that if the horse ever has a choice, it will try to support your commands to it. But still it just doesn't know how to control its own heart rate consciously or ride a giant horse-sized bicycle. Then one day it learns these skills,

and can let you adjust more of its actions. You had all the control the whole time, but all became more.

Consequences

One issue with this concept of taking over is that it isn't clear what it means to 'support' a change. Each change has a number of consequences, and some of them are the point while others are undesirable side effects, such that averting them is an integral part of supporting the change. For instance, moving legs faster means using up blood oxygen and also traveling faster. If you gee up the horse, you want it to support this by replacing the missing blood oxygen, but not to jump on a treadmill to offset the faster travel.

For the horse to get this right in general, it seems that it needs to know about your higher level goals. In practice with horses, they are just built so that if they decide to run faster their respiratory system supplies more oxygen and they aren't struck by a compulsion to get on a treadmill, and if that weren't true we would look for a different animal to ride. The fact that they always assume one kind of thing is the goal of our intervention is fine, because in practice we do basically always want legs for motion and never for using up oxygen.

Maybe there is a systematic difference between desirable consequences and ones that should be offset—in the examples that I briefly think of, the desirable consequences seem more often to do with relationships with larger scale things, and the ones that need offsetting are to do with internal things, but that isn't always true (I might travel because I want to be healthier, but I want to be in the same relationship with those who send me mail). If the situation seems to turn inputs into outputs, then the outputs are often the point, though that is also not always true (e.g. a garbage burner seeks to get rid of garbage, not create smoke). Both of these also seem maybe contingent on our world, whereas I'm interested in a general concept.

Total takeover

I'll set that aside, and for now define a desirable model of controlling a system as something like: the system behaves as it would, but you can adjust aspects of the system and have it support your adjustment, such that the adjustment forwards your goals.

There isn't a clear notion of 'all the control', since at any point there will be things that you can't adjust (e.g. currently the shape of the horse's mitochondria, for a long time the relationship between space and time in the horse system), either because you or the system don't have a means of making the adjustment intentionally, or the system can't support the adjustment usefully. However 'all of the control that anyone has' seems more straightforward, at least if we define who is counted in 'anyone'. (If you can't control the viral spread, is the virus a someone who has some of the universe's control?)

I think whether having all of the control at a particular time gets at what I usually mean by having 'taken over' depends on what we expect to happen with new avenues of control that appear. If they automatically go to whoever had control, then having all of the control at one time seems like having taken over. If they get distributed more randomly (e.g. the horse learns to ride a bicycle, but keeps that power for itself, or a new agent is created with a power), so that your fraction of control deteriorates over time, that seems less like having taken over. If that is how our world is, I think I want to say that one cannot take it over.

***

This was a lot of abstract reasoning. I especially welcome correction from someone who feels they have successfully controlled a horse to a non-negligible degree.

Will transparency help catch deception? Perhaps not

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

In the last few months, some AI safety researchers have signaled what Wei Dai called a "wave of optimism" for using transparency tools to catch deceptive behavior in our machine learning models. By a transparency tool, I mean any technique that allows us to peek inside a model to get a better idea of what goes on inside of it. This post outlines one reason to be pessimistic about the usefulness of using transparency tools to catch advanced deception.

First, I construct a model for why we need transparency tools in the first place. I imagine the training process as a game played by two adversaries: a deceptive model, and an overseer trying to catch any deceptive reasoning. The issue for the overseer is that the deceptive model is opaque, which means that the overseer must use some tools to interpret its thoughts and behavior.

While by design, the deceptive model and overseer are comparably intelligent, the nature of using transparency tools enables a "weakest link" dynamic, implying that the deceptive model can exploit weaknesses in the transparency procedure to fool the overseer. In particular, deception can be hidden if the deceptive model has a good understanding of the deployment environment that isn't revealed while using transparency tools.

In the end, I consider whether this weakness implies that transparency tools act as unnecessary middlemen for catching deception, and use an analogy to GANs to argue that training an overseer to detect deception end-to-end might work better.

The model

Our goal is to take a neural network and detect whether there are any inputs to the network that will cause it to act unacceptably. By assumption, we are concerned about deceptive alignment, a scenario where a misaligned model understands its training process, and will therefore act aligned during training while plotting something else for later.

Since we want to detect deceptive thoughts, we are interested in opening up the model and seeing what goes on inside. We have access to an overseer that interprets the model and spends time thinking whether the model has inputs which cause it to act catastrophically, which will allow us to penalize the model accordingly. Crucially, this overseer is not able to understand the model weights by themselves: instead it relies on a transparency procedure that converts the model into a relatable abstraction, which it then uses to determine whether the model is deceptive.

Formally, we can imagine the dynamic as the following. We are training a model, m

which comes from a class of models M. A transparency function maps models in M to

an abstract description of what the model does, $T : M \to D$. Descriptions can be construed very broadly, and they can include natural language descriptions of the model's decision process, charts and graphs, or even pseudocode. This description is fed to an overseer $O : D \to \{\text{True}, \text{False}\}$, which detects whether the description implies that the model will act deceptively in the deployment environment.

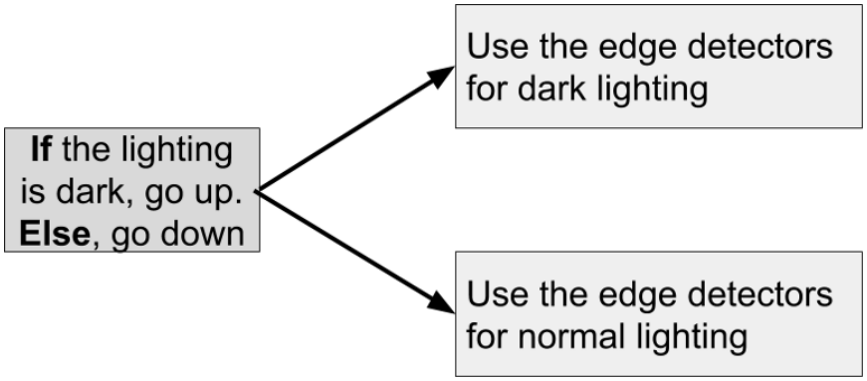
We can think of the transparency function as making a tradeoff between two variables: brevity of the description, and faithfulness of the description to the model m . If the description is too long, or complex, then there is no point of using a transparency function in the first place, since it doesn't actually make the model any more transparent. On the other hand, if the transparency function produces descriptions that are too short, then another problem occurs: the overseer will miss information that allows it to deduce the deceptive capabilities of m.

Achieving a good balance of these variables will determine whether we should use transparency tools in the first place. A good transparency scheme will ideally provide a concise description of how the model makes decisions, while ensuring that nothing important is left out. We will now explore an example of this dynamic in an image classifier, and I will suggest that striking a perfect balance could be difficult.

A deceptive image classifier

Suppose we are training a neural network to recognize cars using some advanced neural network design. This neural network is smarter than your average convolutional neural network. In particular, rather than imagining it only as a set of neural layers with edge detectors, we also imagine that it has adaptive computational layers that can change depending on the gradient updates. At the extreme end, the training process could even nudge the neural network into arbitrary computational graphs, which allows for if-statements, explicit memory, and loops.

Allowing the model to adapt its structure during the gradient updates is very important, as it enables the model to be more robust to different domain distributions. For illustration of this effect, consider if the gradient updates produce the following schematic for the model.



This design allows the model to be robust to different type of lighting, which demonstrates the utility of the adaptive approach. However, there's also a dark side to this adaptivity. While normally we think of models having only a choice over their input/output behavior, a sufficiently competent model could also understand the whole training procedure, which means that it could use gradient hacking to route a path through the space of computational graphs.

Since the deceptive model can ask questions about what setting they are in, they can exploit this functionality by inserting an if-statement that checks whether they are currently inside the training environment. During training, we can expect that the model will act safely. During deployment, however, the switch will be activated and the model will pursue its own devious plans.

That's where the transparency tools and overseer come in.

The transparency function has the job of creating a relatable description of the model that would allow the overseer to catch any attempted deception. Of course, the transparency function can't just rely on using the input/output behavior of the model during training to produce a faithful description. Doing this would create a misleading impression, as it wouldn't allow the overseer to detect the dangers involved in a distribution shift.

Instead, the transparency function must convert the pile of matrices, and thus the internal structure of the model, into a relatable abstraction. But now this allows for the following troubling dynamic.

If the deceptive model understands how the transparency tools work, then it can modify its own architecture so that the transparency function leaves out important information. In the above example, the deceptive model might realize that the deployment environment will occur at night, and therefore use this if-statement as its signal to perform a treacherous turn. Further, since the deceptive model understands the transparency tool, it knows how to place if-statements that will be glossed over.

The transparency tools are intended to highlight decisions like this, but by assumption, the transparency tools are less 'intelligent' than the deceptive model. Therefore, it may choose to omit this if-statement in its description of the model, and focus on information it believes is more critical.

In other words, not realizing that the nighttime-check was the signal for a treacherous turn, the transparency tool glosses over this if-statement, and doesn't provide it to the overseer.

This ends in catastrophe once the deception is successful. The overseer *is* comparably intelligent to the deceptive model, but since the if-statement above was omitted in the transparency description, the overseer was not able to deduce the dangerous off-distribution behavior.

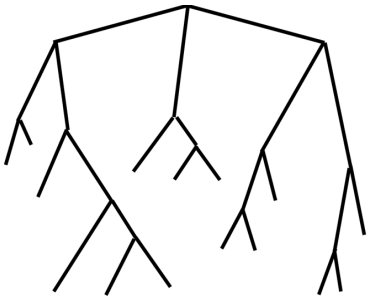
Therefore, the deceptive model successfully fooled the overseer, even though both players in the game were comparably intelligent.

Why are we assuming that the transparency tools are less intelligent than the deceptive

model?

An immediate objection is simply, "Yes but a good transparency tool would not have omitted the if-statement." Ultimately, I think this is a weak objection.

The first thing to note is that is that the transparency function wasn't totally dumb. Understanding the nature of the deployment environment is hard. The transparency function was just doing its job by trading off brevity and faithfulness. To make this point more compelling, we can consider that the true computational graph was really a large decision tree, with a huge branching factor for the various types of situations the model could be in.



In my example, the transparency tool is really just doing its best job of pruning the irrelevant parts of the tree. If you didn't know the important detail that the deployment environment was at night, then checking whether there is dark or dim lighting could simply be seen as a side feature that allowed the model to be more robust, and something like that can be omitted.

More importantly, if we allow the transparency tools to be more intelligent than the deceptive model, we must question the role that the overseer is playing in this dynamic. For instance, if the transparency model was exceptionally competent, then all it needs to do in its description is either say, "This model is deceptive" or say, "This model is not deceptive." Then, the overseer's job is trivial.

This suggests that decomposing safety-checking into two components: a transparency tool, and an overseer that interprets the results of the tool, is a *bad model*. If the transparency tools are too weak, then the deceptive model will exploit this weakness to fool the overseer, even if the overseer is very intelligent. On the other hand, if the tools are very strong, then the overseer plays no role at all -- suggesting that there was never any need to create interpretability tools, since a deception-checker trained end-to-end would have worked fine.

To better highlight why I think this casts doubt on transparency tools, I offer an analogy with GANs.

A fictional suggestion for training GANs

GANs are trained via an adversarial game between two neural networks. One neural network -- the generator -- is trained to produce a counterfeit sample when it is given

some latent noise, and the other -- the discriminator -- is trained to detect whether a sample is counterfeit or from the real distribution.

Suppose someone were to propose the following way to help the discriminator: rather than training it end-to-end to help detect counterfeit samples, we instead make its input equal to the output of a transparency tool that gives it insight into how the current sample looks. The transparency tool could, for instance, highlight if it detects any weird shadows or bad lighting, and then feed this information to the discriminator to produce a final determination.

"Since the transparency tool focuses on the weaknesses of the generator, this will allow the discriminator to more easily detect counterfeit samples," the person argues.

But what the person failed to consider was that all this tool did was *relax* the job of the generator, since now all it needs to do is get past the weird-shadow and bad-lighting detector stage. It doesn't even matter how competent the discriminator is, since it's only given information from the transparency tool.

The person replies, "OK fine, we can give the discriminator the output from the tool, *and* the original sample. That way it can use the tool as a bonus to help it make the determination."

But now, the discriminator is operating over a larger dimension input-space than it would have otherwise, which makes its job harder. If it was simply trained end-to-end, then it could have developed a weird-shadow detector and bad-lighting detector *on its own,* if that were appropriate. But now it has extraneous information, which may or may not be useful for actually detecting counterfeit samples.

Conclusion

Contra Betteridge's law of headlines, I still think that transparency tools can be useful for catching deception. Also, this argument does not imply that transparency tools are useless for tasks other than catching deception. The biggest weakness with my argument is that the alternative I offered was training an overseer end-to-end, which is not possible if we are assuming a human is the overseer.

However, I don't think that this weakness is fatal to my argument: in order for transparency tools to work well, they must be strong enough that their weaknesses can't be exploited by a deceptive model. But we just argued above that if the tools were that strong, they would already be indistinguishable from a deception checker trained end-to-end.

Finally, I suspect people will object to my argument by saying that I am taking a very narrow approach to transparency tools, and that a model trained end-to-end to detect deception *still counts* as a transparency tool. I have little to say about this objection other than, I'm not arguing against *that type* of tool.

Antimemes

Antimemes are self-keeping secrets. You can only perceive an antimeme if you already know it's there. Antimemes don't need a conspiracy to stay hidden because you can't comprehend an antimeme just by being told it exists. You can shout them to the heavens and nobody will listen. I'll try to explain with a fictitious example.

Suppose we all had an invisible organ behind our ears and our brains kept it secret from our consciousness. If I told you "you have an invisible organ behind your ear" you wouldn't believe me. You'd only believe it exists if you deduced its existence from a trail of evidence.

You can deduce the existence of an antimeme from the outline of the hole it cuts in reality. If you find an old photo with a gap where a person has been painted out then you can be confident that someone has been disappeared. You can then figure out who it is with conventional investigative methods. The challenge is noticing the gap in the first place and then not dismissing it as noise.

Different cultures have different antimemes. The more different two cultures are from each other the less their antimemes overlap. You can sweep up a mountain of antimemes just by reading a Chinese or Arabic history of civilization and comparing it to Western world history. You can snag a different set by learning what it was like to live in a hunter-gatherer or pastoralist society.

You can do the same thing with technology. Developing a proficiency in Lisp will shatter your tolerance of inferior programming languages. Once you've internalized `defmacro` you can never go back.

As for jobs: once an entrepreneur, always an entrepreneur[1].

Comprehending an antimeme takes work. You slog toward it for a long time and then eventually something clicks like a ratchet. Until then everything you've learned is reversible. After it clicks you've permanently unlocked a new level of experience, like stream entry.

Stream entry is another antimeme, by the way.

Antimemes are easily dismissed as pseudoscience. Pseudoscience is a meme, not an antimeme. You can distinguish antimemes from pseudoscience at a glance by examining why they're suppressed. Pseudoscience is dismissed as fraudulent. Antimemes are dismissed as inapposite.

---

1. There are two different kinds of entrepreneurship. The more common form of entrepreneurship is self-employment where you sell your labor. I'm not talking about this common entrepreneurship. Entrepreneurship where you exploit an overlooked market opportunity is an antimeme. ↩

Explaining why false ideas spread is more fun than why true ones do

As typical for a discussion of memes (of the Richard Dawkins variety), I'm about to talk about something completely unoriginal to me, but that I've modified to some degree after thinking about it.

The thesis is this: there's a tendency for people to have more interest in explaining the spread of ideas they think are false, when compared to ideas they think are true.

For instance, there's a lot written about how and why religion spread through the world. On the other hand, there's comparatively little written about how and why general relativity spread through the world. But this is strange -- they are both just ideas that are spread via regular communication channels.

One could say that the difference is that general relativity permits experimental verification, and therefore it's no surprise that it spread through the world. The standard story here is that since the idea is simply true, the explanation for why it became widespread is *boring* -- people merely became convinced due to its actual veracity.

I reject this line of thought for two reasons. First, the vast majority of people don't experimentally verify general relativity, or examine its philosophical basis. Therefore, the mechanism by which the theory spreads is probably fairly similar to religion. Secondly, I don't see why the idea being true makes the memetic history of the idea any less interesting.

I'm not really sure about the best explanation for this effect -- that people treat true memes as less interesting than false ones -- but I'd like to take a guess. It's possible that the human brain seeks simple *single* stories to explain phenomena, even if the real explanation for those phenomena are due to a large number of factors. Furthermore, humans are bored by reality: if something has a seemingly clear explanation, even if the speaker doesn't *actually* know the true explanation, it's nonetheless not very fun to speculate about.

This theory would predict that we would be less interested in explaining why true memes spread, because we already have a readily available story for that: namely, that the idea is true and therefore compels its listeners to believe in it. On the other hand, a false meme no longer permits this standard story, which forces us to search for an alternative, perhaps exciting, explanation.

One possible takeaway is that we are just extremely wrong about why some ideas spread through the world. It's hard enough to know why a single person believes what they do. The idea that a single story could adequately explain why *everyone* believes something is even more ludicrous.

Robin Hanson on the futurist focus on AI


Robin Hanson

Robert Long and I recently talked to Robin Hanson—GMU economist, prolific blogger, and longtime thinker on the future of AI—about the amount of futurist effort going into thinking about AI risk.

It was noteworthy to me that Robin thinks human-level AI is a century, perhaps multiple centuries away— much longer than the 50-year number given by AI researchers. I think these longer timelines are the source of a lot of his disagreement with the AI risk community about how much of futurist thought should be put into AI.

Robin is particularly interested in the notion of 'lumpiness'– how much AI is likely to be furthered by a few big improvements as opposed to a slow and steady trickle of progress. If, as Robin believes, most academic progress and AI in particular are not likely to be 'lumpy', he thinks we shouldn't think things will happen without a lot of warning.

The full recording and transcript of our conversation can be found here.

Relevance Norms; Or, Gricean Implicature Queers the Decoupling/Contextualizing Binary

Reply to: Decoupling vs Contextualising Norms

Chris Leong, following John Nerst, distinguishes between two alleged discursive norm-sets. Under "decoupling norms", it is understood that claims should be considered in isolation; under "contextualizing norms", it is understood that those making claims should also address potential implications of those claims in context.

I argue that, at best, this is a false dichotomy that fails to clarify the underlying issues—and at worst (through no fault of Leong or Nerst), the concept of "contextualizing norms" has the potential to legitimize derailing discussions for arbitrary political reasons by eliding the key question of *which* contextual concerns are *genuinely relevant*, thereby conflating legitimate and illegitimate bids for contextualization.

Real discussions adhere to what we might call "relevance norms": it is almost universally "eminently reasonable to expect certain contextual factors or implications to be addressed." Disputes arise over *which* certain contextual factors those are, not *whether* context matters at all.

The standard academic account explaining how what a speaker means differs from what the *sentence* the speaker said means, is H. P. Grice's theory of conversational implicature. Participants in a conversation are expected to add neither more nor less information than is needed to make a *relevant* contribution to the discussion.

Examples abound. If I say, "I ate some of the cookies", I'm *implicating* that I didn't eat *all* of the cookies, because if I had, you would have expected me to say "all", not "some" (even though the decontextualized sentence "I ate some of the cookies" is, in fact, true).

Or suppose you're a guest at my house, and you ask where the washing machine is, and I say it's by the stairs. If the machine then turns out to be broken, and you ask, "Hey, did you know your washing machine is broken?" and I say, "Yes", you're probably going to be pretty baffled why I didn't say "It's by the stairs, *but you can't use it because it's broken*" earlier (even though the decontextualized answer "It's by the stairs" was, in fact, true).

Leong writes:

> Let's suppose that blue-eyed people commit murders at twice the rate of the rest of the population. With decoupling norms, it would be considered churlish to object to such direct statements of facts. With contextualising norms, this is deserving of criticism as it risks creates a stigma around blue-eyed people.

With relevance norms, objecting might or might not make sense depending on the context in which the direct statement of fact is brought up.

Suppose Della says to her Aunt Judith, "I'm so excited for my third date with my new boyfriend. He has the most beautiful blue eyes!"

Judith says, "Are you sure you want to go out with this man? Blue-eyed people commit murders at twice the rate of the general population."

How should Della reply to this? Judith is just in the wrong here—but *not* as a matter of a subjective choice between "contextualizing" and "decoupling" norms, and not because blue-eyed people are a sympathetic group who we wish to be seen as allied with and don't want to stigmatize. Rather, the probability of getting murdered on a date is quite low, *and* Della already has a lot of individuating information about whether her boyfriend is likely to be a murderer from the previous two dates. Maybe (Fermi spitballing here) the evidence of the boyfriend's eye color raises Della's probability of being murdered from one-in-a-million to one-in-500,000? Judith's bringing the possibility up *at all* is a waste of fear in the same sense that lotteries are said to be a waste of hope. Fearmongering about things that are almost certainly not going to happen is *uncooperative*, in Grice's sense—just like it's uncooperative to tell people where to find a washing machine that doesn't work.

On the other hand, if I'm making a documentary film interviewing murderers in prison and someone asks me why so many of my interviewees have blue eyes, "Blue-eyed people commit murders at twice the rate of the rest of the population" is a *completely relevant reply*. It's not clear how else I could possibly answer the question without making reference to that fact!

So far, *relevance* has been a black box in this exposition: unfortunately, I don't have an elegant reduction that explains what *cognitive algorithm* makes some facts seem "relevant" to a given discussion. But hopefully, it should now be intuitive that the determination of what context is relevant is the consideration that is, um, relevant. Framing the matter as "decouplers" (context doesn't matter!) *vs*. "contextualizers" (context matters!) is misleading because once "contextualizing norms" have been judged admissible, it becomes easy for people to motivatedly derail any discussions they don't like with endless isolated demands for contextualizing disclaimers.

Matthew Walker's "Why We Sleep" Is Riddled with Scientific and Factual Errors

This is a linkpost for https://guzey.com/books/why-we-sleep/#the-full-discussion-of-sleep-deprivation-therapy-from-chapter-7

This was one of the most thought-provoking posts I read this month. Mostly because I spent a really large number of hours of my life sleeping, and also significantly increased the amount that I've been sleeping over the past three years, and this has me seriously considering reducing that number again.

The opening section of the article:

> Matthew Walker (a) is a professor of neuroscience and psychology at the University of California, Berkeley, where he also leads the Center for Human Sleep Science.
>
> His book *Why We Sleep* (a) was published in September 2017. Part survey of sleep research, part self-help book, it was praised by The New York Times (a), The Guardian (a), and many others. It was named one of NPR's favorite books of 2017. After publishing the book, Walker gave a TED talk, a talk at Google, and appeared on Joe Rogan's and Peter Attia's podcasts. A month after the book's publication, he became (a) a sleep scientist at Google.
>
> On page 8 of the book, Walker writes:
>
> **> [T]he real evidence** that makes clear all of the dangers that befall individuals and societies when sleep becomes short have not been clearly telegraphed to the public … In response, this book is intended to serve as a **scientifically accurate intervention** addressing this unmet need **[emphasis in this quote and in all quotes below mine]**
>
> In the process of reading the book and encountering some extraordinary claims about sleep, I decided to compare the facts it presented with the scientific literature. I found that the book consistently overstates the problem of lack of sleep, sometimes egregiously so. It misrepresents basic sleep research and contradicts its own sources.
>
> In one instance, Walker claims that sleeping less than six or seven hours a night doubles one's risk of cancer – this is not supported by the scientific evidence. In another instance, **Walker seems to have invented a "fact" that the WHO has declared a sleep loss epidemic.** In yet another instance, he falsely claims that the National Sleep Foundation recommends 8 hours of sleep per night, and then uses this "fact" to falsely claim that two-thirds of people in developed nations sleep less than the "the recommended eight hours of nightly sleep" – a myth that spread like wildfire after the book's publication.

Walker’s book has likely wasted thousands of hours of life and worsened the health of people who read it and took its recommendations at face value.

Any book of *Why We Sleep’s* length is bound to contain some factual errors. Therefore, to avoid potential concerns about cherry-picking the few inaccuracies scattered throughout, **in this essay, I’m going to highlight the five most egregious scientific and factual errors Walker makes *in Chapter 1* of the book**. This chapter contains 10 pages and constitutes less than 4% of the book by the total word count.

The new dot com bubble is here: it’s called online advertising

This is a linkpost for https://thecorrespondent.com/100/the-new-dot-com-bubble-is-here-its-called-online-advertising/13228924500-22d5fd24

If you want to understanding Goodharting in advertising, this is a great article for that.

At the heart of the problems in online advertising is selection effects, which the article explains with this cute example:

> Picture this. Luigi’s Pizzeria hires three teenagers to hand out coupons to passersby. After a few weeks of flyering, one of the three turns out to be a marketing genius. Customers keep showing up with coupons distributed by this particular kid. The other two can’t make any sense of it: how does he do it? When they ask him, he explains: "I stand in the waiting area of the pizzeria."
>
> It’s plain to see that junior’s no marketing whiz. Pizzerias do not attract more customers by giving coupons to people already planning to order a quattro stagioni five minutes from now.

The article goes through an extended case study at eBay, where selection effects were causing particularly expensive results without anyone realizing it for years:

> The experiment continued for another eight weeks. What was the effect of pulling the ads? Almost none. For every dollar eBay spent on search advertising, they lost roughly 63 cents, according to Tadelis’s calculations.
>
> The experiment ended up showing that, for years, eBay had been spending millions of dollars on fruitless online advertising excess, and that the joke had been entirely on the company.
>
> To the marketing department everything had been going brilliantly. The high-paid consultants had believed that the campaigns that incurred the biggest losses were the most profitable: they saw brand keyword advertising not as a $20m expense, but a $245.6m return.

The problem, of course, is Goodharting, by trying to optimize for something that's easy to measure rather than what is actually cared about:

> The benchmarks that advertising companies use – intended to measure the number of clicks, sales and downloads that occur after an ad is viewed – are fundamentally misleading. None of these benchmarks distinguish between the selection effect (clicks, purchases and downloads that are happening anyway) and the advertising effect (clicks, purchases and downloads that would not have happened without ads).

And unsurprisingly, there's an alignment problem hidden in there:

> It might sound crazy, but companies are not equipped to assess whether their ad spending actually makes money. It is in the best interest of a firm like eBay to

> know whether its campaigns are profitable, but not so for eBay’s marketing department.
>
> Its own interest is in securing the largest possible budget, which is much easier if you can demonstrate that what you do actually works. Within the marketing department, TV, print and digital compete with each other to show who’s more important, a dynamic that hardly promotes honest reporting.
>
> The fact that management often has no idea how to interpret the numbers is not helpful either. The highest numbers win.

To this I'll just add that this problem is somewhat solvable, but it's tricky. I previously worked at a company where our entire business model revolved around calculating lift in online advertising spend by matching up online ad activity with offline purchase data, and a lot of that involved having a large and reliable control group against which to calculate lift. The bad news, as we discovered, was that the data was often statistically underpowered and could only distinguish between negative, neutral, and positive lift and could only see not neutral lift in cases where the evidence was strong enough you could have eyeballed it anyway. And the worse news was that we had to tell people their ads were not working or, worse yet, were lifting the performance of competitor's products.

Some marketers' reactions to this were pretty much as the authors' capture it:

> Leaning on the table, hands folded, he gazed at his hosts and told them: "You’re fucking with the magic."

AI Alignment Research Overview (by Jacob Steinhardt)

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://docs.google.com/document/d/1FbTuRvC4TFWzGYerTKpBU7FJlyvjeOvVYF2uYNFSlOc/edit

I'm really excited to see someone outline all the work they think needs solving in AI alignment - to describe what the problem looks like, what a solution looks like, and what work has been done so far. Especially from Jacob, who is a coauthor of the Concrete Problems in AI Safety paper.

Below, I've included some excerpts from doc. I've included the introduction, the following section describing the categories of technical work, and some high-level information from the long sections on 'technical alignment problem' and the 'detecting failures in advance'.

Introduction

This document gives an overview of different areas of technical work that seem necessary, or at least desirable, for creating safe and aligned AI systems. The focus is on safety and alignment of powerful AI systems, i.e. systems that may exceed human capabilities in a broad variety of domains, and which likely act on a large scale. Correspondingly, there is an emphasis on approaches that seem scalable to such systems.

By “aligned”, I mean that the actions it pursues move the world towards states that humans want, and away from states that humans don’t want. Some issues with this definition are that different humans might have different preferences (I will mostly ignore this issue), and that there are differences between stated preferences, “revealed” preferences as implied by actions, and preferences that one endorses upon reflection (I won’t ignore this issue).

I think it is quite plausible that some topics are missing, and I welcome comments to that regard. My goal is to outline a critical mass of topics in enough detail that someone with knowledge of ML and some limited familiarity with AI alignment as an area would have a collection of promising research directions, a mechanistic understanding of why they are promising, and some pointers for what work on them might look like.

To that end, below I outline four broad categories of technical work: **technical alignment** (the overcoming of conceptual or engineering issues needed to create aligned AI), **detecting failures** (the development of tools for proactively assessing the safety/alignment of a system or approach), **methodological understanding** (best practices backed up by experience), and **system-building** (how to tie together the three preceding categories in the context of many engineers working on a large system). These are described in more detail in the next section.

In each section I give examples of problems we might want to solve. I imagine these in the context of future powerful AI systems, which means that most of the concrete scenarios are speculative, vague, and likely incorrect if interpreted as a prediction about the future. If I were to give the strongest justification for the research topics below, I would instead focus on near-future and existing systems, which already exhibit many of the issues I discuss. Nevertheless, I think this imaginative exercise can be helpful both for stimulating research and for keeping the focus on scalable solutions.

Caveats. I found it difficult to write a research overview of a field as nascent as AI alignment, as anything I could write sounded either too authoritative relative to my confidence, or so full of caveats and qualifications as to be unreadable. I settled for eliding many of the qualifications and providing this single caveat up front: that this document reflects an imperfect snapshot of my current thinking, that it expresses many ideas more sloppily than I would usually feel comfortable putting into writing, and that I hope readers will forgive this sloppiness in the service of saying *something* about a topic that I feel is important.

This document is not meant to be a description of *my personal interests*, but rather of potentially promising topics within a field I care about. My own interests are neither a subset nor superset of the topics in this document, although there is high overlap. Even confined to AI alignment, this document is out-of-date and omits some of my recent thinking on economic aspects of ML.

Finally, I make a number of claims below about what research directions I think are promising or un-promising. Some of these claims are likely wrong, and I could even imagine changing my mind after 1 hour of conversation with the right person. I decided that this document would be more informative and readable if I gave my unfiltered take (rather than only opinions I thought I would likely defend upon consideration), but the flip side is that if you think I'm wrong about something, you should let me know!

Categories of technical work

In this document, I will discuss four broad categories of technical work:

Technical alignment problem. Research on the "technical alignment problem" either addresses conceptual obstacles to making AI aligned with humans (e.g. robustness, reward mis-specification), or creates tools and frameworks that aid in making AI aligned (e.g. scalable reward generation).

Detecting failures in advance. Independently of having solved various alignment problems, we would like to have ways of probing systems / blueprints of systems to know whether they are likely to be safe. Example topics include interpretability, red-teaming, or accumulating checklists of failure modes to watch out for.

Methodological understanding. There is relatively little agreement or first-hand knowledge of how to make systems aligned or safe, and even less about which methods for doing so will scale to very powerful AI systems. I am personally skeptical of our ability to get alignment right based on purely abstract arguments without also having a lot of methodological experience, which is why I think work in this category is important. An example of a methodology-focused document is Martin Zinkevich's Rules of Reliable ML, which addresses reliability of existing large systems.

System-building. It is possible that building powerful AI will involve a large engineering effort (say, 100+ engineers, 300k+ lines of code). In this case we need a framework for putting many components together in a safe way.

Technical alignment problem

We would ideally like to build AI that acts according to some specification of human values, and that is robust both to errors in the specification and to events in the world. To achieve this robustness, the system likely needs to represent uncertainty about both its understanding of human values and its beliefs about the world, and to act appropriately in the face of this uncertainty to avoid any catastrophic events.

I split the technical alignment problem correspondingly into four sub-categories:

Scalable reward generation. Powerful AI systems will potentially have to make decisions in situations that are foreign to humans or otherwise difficult to evaluate---for instance, on scales far outside human experience, or involving subtle but important downstream consequences. Since modern ML systems are primarily trained through human-labeled training data (or more generally, human-generated reward functions), this presents an obstacle to specifying which decisions are good in these situations. Scalable reward generation seeks to build processes for generating a good reward function.

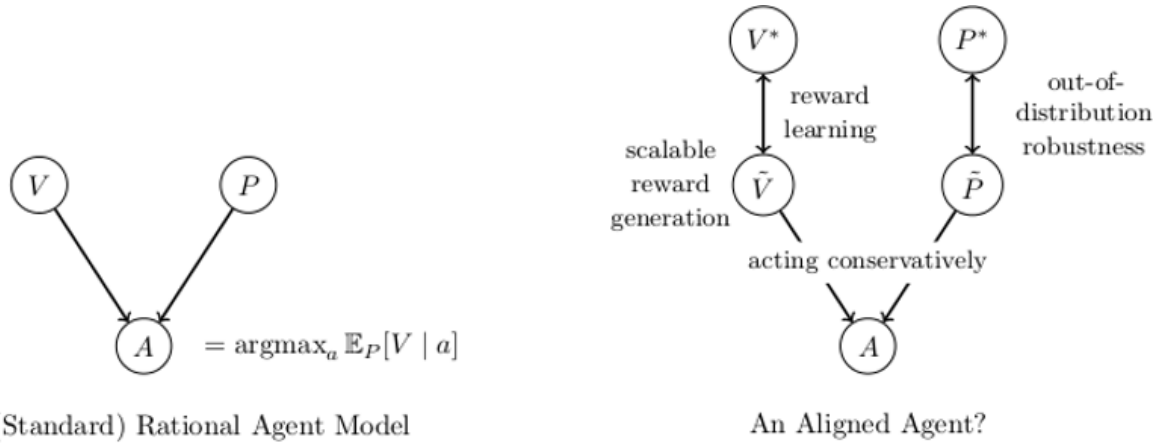
Reward learning. Many autonomous agents seek to maximize the expected value of some reward function (or more broadly, to move towards some specified goal state / set of states). Optimizing against the reward function in this way can cause even slight errors in the reward to lead to large errors in behavior--typically, increased reward will be well-correlated with human-desirability for a while, but will become anti-correlated after a point. Reward learning seeks to reason about differences between the observed (proxy) reward and the true reward, and to converge to the true reward over time.

Out-of-distribution robustness is the problem of getting systems to behave well on inputs that are very different from their training data. This might be done by a combination of transfer learning (so the system works well in a broader variety of situations) and having more uncertainty in the face of unfamiliar/atypical inputs (so the system can at least notice where it is likely to not do well).

Acting conservatively. Safe outcomes are more likely if systems can notice situations where it is unclear how to act, and either avoid encountering them, take actions that reduce the uncertainty, or take actions that are robustly good. This would, for instance, allow us to specify an ambiguous reward function that the system could clarify as needed, rather than having to think about every possible case up-front.

Acting conservatively interfaces with reward learning and out-of-distribution robustness, as the latter two focus on noticing uncertainty while the former focuses on what to do *given* the uncertainty. Unfortunately, current methods for constructing uncertainty estimates seem inadequate to drive such decisions, and even given a good uncertainty estimate little work has been done on how the system should use it to shape its actions.

A toy framework. Conceptually, it may be useful to think in terms of the standard rational agent model, where an agent has a value function or utility function V, and beliefs P, and then takes actions A that maximize the expected value of V under P (conditioned on the action A). Failures of alignment could come from incorrect beliefs P, or a value function V that does not lead to what humans want. Out-of-distribution robustness seeks to avoid or notice problems with P, while scalable reward generation seeks to produce accurate information about some value function V that is aligned with humans. Reward learning seeks to correct for inaccuracies in the reward generation process, as well as the likely limited amount of total data about rewards. Finally, acting conservatively takes into account the additional uncertainty due to acting out-of-distribution and having a learned reward function, and seeks to choose actions in a correspondingly conservative manner.



(Standard) Rational Agent Model

An Aligned Agent?

In an RL setting where we take actions via a learned policy, we can tell the same story but with a slightly modified diagram. Instead of an action A we have a learned policy θ, and instead of $P*$ and $\tilde{P}$ denoting beliefs, they denote distributions over environments ($P*$ is the true on-policy environment at deployment time, while $\tilde{P}$ is the distribution of training environments).

Other topics. Beyond the topics above, the problem of **counterfactual reasoning** cuts across multiple categories and seems worth studying on its own. There may be other important categories of technical work as well.

Detecting failures in advance

The previous section lays out a list of obstacles to AI alignment and technical directions for working on them. This list may not be exhaustive, so we should also develop tools for discovering new potential alignment issues. Even for the existing issues, we would like ways of being more confident that we have solved them and what sub-problems remain.

While machine learning often prefers to hew close to empirical data, much of the roadmap for AI alignment has instead followed from more abstract considerations and thought experiments, such as asking “What would happen if this reward function were optimized as far as possible? Would the outcome be good?” I actually think that ML undervalues this abstract approach and expect it to continue to be fruitful, both for pointing to useful high-level research questions and for analyzing concrete systems and approaches.

At the same time, I am uncomfortable relying solely on abstract arguments for detecting potential failures. Rigorous empirical testing can make us more confident that a problem is actually solved and expose issues we might have missed. Finding concrete instantiations of a problem can both more fruitfully direct work and convince a larger set of people to care about it (as in the case of adversarial examples for images). More broadly, empirical investigations have the potential to reveal new issues that were missed under purely abstract considerations.

Two more empirically-focused ways of detecting failures are **model probing/visualization** and **red-teaming,** discussed below. Also valuable is **examining trends** in ML. For instance, it looks to me like reward hacking in real deployed systems is becoming a bigger issue over time; this provides concrete instances of the problem to examine for insight, gives us a way to measure how well we’re doing at the problem, and helps rally a community around the

problem. Examining trends is also a good way to take an abstract consideration and make it more concrete.

Wrinkles

Why does our skin form wrinkles as we age?

This post will outline the answer in a few steps:

- Under what conditions do materials form wrinkles, in general?
- How does the general theory of wrinkles apply to aging human skin?
- What underlying factors drive the physiological changes which result in wrinkles?

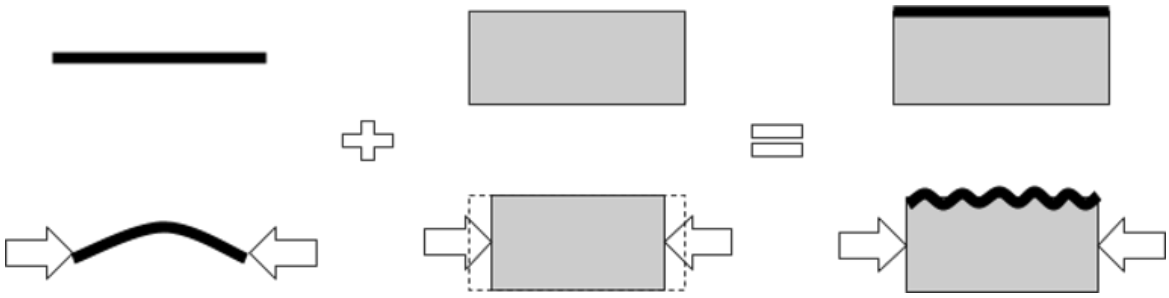
In the process, we'll draw on sources from three different fields: mechanical engineering, animation, and physiology.

Why do Materials Wrinkle?

Imagine we have a material with two layers:

- A thin, stiff top layer
- A thick, elastic bottom layer

We squeeze this material from the sides, so the whole thing compresses.

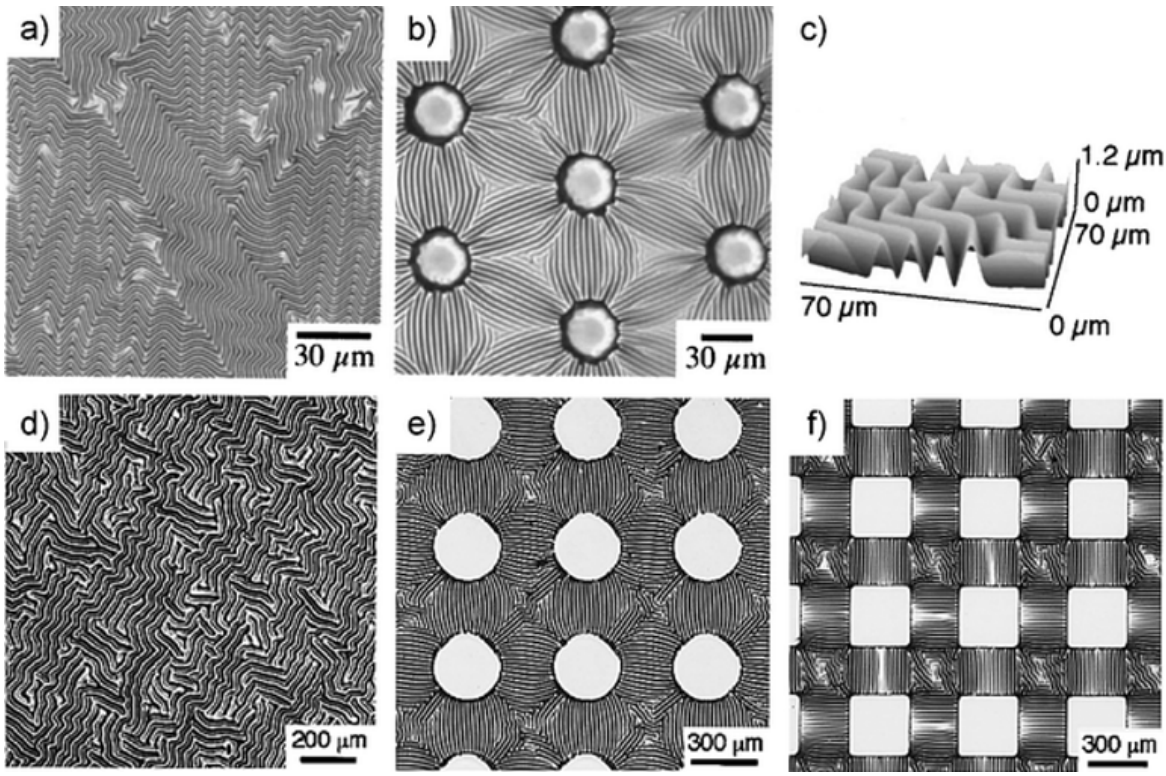


The two layers want to do different things under compression:

- The thin top layer maintains its length but wants to minimize bending, so it wants to bow outward and form an arc
- The elastic bottom layer wants to minimize vertical displacement, so it wants to just compress horizontally without any vertical change at all.

Because the two layers are attached, these two objectives trade off, and the end result is waves - aka wrinkles. Longer waves allow the top layer to bend less, so a stiffer top layer yields longer waves. Shorter waves allow the bottom layer to expand/compress less vertically, so a stiffer bottom layer yields shorter waves. The "objectives" can be quantified via the energy associated with bending the top layer or displacing the bottom layer, leading to quantitative predictions of the wavelength - see this great review paper for the math.

Engineers do this with a thin metal coating on soft plastic. The two are bound together at high temperature, and then the whole system compresses as it cools. The end result is cool wrinkle patterns:



Other interesting applications include predicting mountain spacing (with crust and mantle as the two layers) and surface texture of dried fruit - see the review paper for more info and cool pictures.

The same thing happens in skin.

Skin Layers

For our purposes, skin has three main layers:

- The epidermis is a thin, relatively stiff top layer
- The SENEB (subepidermal non-echogenic band, also sometimes called subepidermal low-echogenic band, SLEB) is a mysterious age-related layer, mostly absent in youth and growing with age, between the epidermis and dermis - more on this later
- The dermis is the thick base layer, containing all the support structure - blood vessels, connective tissue, etc

Both the SENEB and the dermis are relatively thick, elastic layers, while the epidermis is thin and stiff. So, based on the model from the previous section, we'd expect this system to form wrinkles.

But wait, if our skin has a thin stiff top layer and thick elastic bottom layer even in youth, then why do wrinkles only form when we get old?

Turns out, young people have wrinkles too. In youth, the wrinkles have short wavelength - we have lots of tiny wrinkles, so they're not very visible. As we age, our wrinkle-wavelength grows, so we have fewer, larger wrinkles - which are more visible. The real question is not "why do wrinkles form as we age?" but rather "why does the wavelength of wrinkles grow as we age?".

Based on the simple two-layer model, we'd expect that either the epidermis becomes more stiff with age, or the lower layers become less stiff.

This the right basic idea, but of course it's a bit more complicated in practice. These guys use a three-layer model, cross-reference parameters from the literature with what actually reproduces realistic age-related wrinkling (specifically for SENEB modulus), and find realistic age-related wrinkles with these numbers:

	Thickness	Elastic Modulus
Epidermis	15 -> 15 um	6 ->12 MPa
SENEB	50 -> 200 um	.05 -> .05 MPa
Dermis	1.2 -> 1.1 mm	.6 -> 1.0 MPa

(arrows indicate change from young to old). Other than the SENEB elastic modulus, all of these numbers are derived from empirically measured parameters - see the paper for details.

Age-Related Physiological Changes

We have two main questions left:

- Why do the dermis and epidermis stiffen with age?
- What exactly is the SENEB, and why does it grow with age?

I haven't looked too much into stiffening of the dermis, but the obvious hypothesis is that it stiffens for the same reason lots of other tissues stiffen with age. At some point I'll have a post on stiffening of the vasculature which will talk about that in more depth, but for now I'm going to punt.

The paper from the previous section notes that the epidermis stiffens mainly due to dehydration; rehydrating the epidermis reverses the stiffening (this is the basis of many cosmetics). A dehydrated epidermis makes sense, since both the SENEB and age-related problems in the vasculature will isolate the epidermis more from the bloodstream (although I haven't seen direct experimental evidence of that causal link).

That leaves the mysterious SENEB. What is it, and why does it grow with age?

The name "subepidermal non-echogenic band" is a fancy way of saying that there's a layer under the epidermis which is transparent to ultrasound imaging. That's the main way the SENEB is detected: it shows up as a space between the epidermis and dermis on ultrasound images of the skin.

As far as I can tell, little is known about the SENEB. The main things we do know:

- SENEB grows with age; see numbers above
- SENEB is found in aged skin typically exposed to sunlight ("photoaged", e.g. hands and face) but not in hidden skin (e.g. butt).

Most authors claim that the SENEB consists of elastin deposits. That matches what we know of solar elastosis, the build-up of elastin deposits in photoaged skin. But I haven't seen anyone systemically line up the ultrasonic and histologic images and chemically analyze the

SENEB layer to check that it really is made of elastin. (This may just be a case of different researchers with different tools using different names for things which are the same.)

Assuming that the SENEB does consist of accumulated elastin, why is elastin accumulating? Well, it turns out that elastin is never broken down in humans. It does not turn over. On the other hand, the skin presumably needs to produce new elastin sometimes to heal wounds. Indeed, many authors note that the skin's response to UV exposure is basically a wound-healing response. Again, I haven't seen really convincing data, but I haven't dug too thoroughly. It's certainly plausible that elastin is produced in response to UV as part of a wound-healing response, and then accumulates with age. That would explain why the SENEB grows in photoaged skin, but not in hidden skin.

Pieces of time

My friend used to have two ‘days’ each day, with a nap between—in the afternoon, he would get up and plan his day with optimism, whatever happened a few hours before washed away. Another friend recently suggested to me thinking of the whole of your life as one long day, with death on the agenda very late this evening. I used to worry, when I was very young, that if I didn’t sleep, I would get stuck in yesterday forever, while everyone else moved on to the new day. Right now, indeed some people have moved on to Monday, but I’m still winding down Sunday because I had a bad headache and couldn’t sleep. Which is all to say, a ‘day’ does not just mean a 24 hour measure of time, in our minds. Among its further significance, we treat it as a modular unit: we expect things within it to be more continuous and intermingled with each other than they are with things outside of it. What happens later today is more of a going concern at present than something that happens after sleeping. The events of this morning are more part of a continuous chapter, expected to flavor the present, than what happened yesterday. The same is true to some extent for weeks, months and years (but not for fortnights or periods of 105 hours).

I think days are well treated as modular like this because sleeping really separates them in relevant ways. I notice two other kinds of natural modular time-chunks that seem worth thinking in terms of, but which I don’t have good names for:

- Periods during which you are in one context and stream of thought (usually a minute to a few hours long). For instance the period of going for a walk, or the period between getting home and receiving a phone call that throws you into a new context and set of thoughts. During one such chunk, I can remember a lot about the series of thoughts so far, and build upon them. Whereas if I try to go back to them later, they are hard to bring back to life, especially the whole set of thoughts and feelings that I wandered around during a period, rather than just a single insight brought from it. Within chunks like this, my experience seems more continuous and intermingled with other experience within the chunk. Then I get an engaging message or decide to go out, and a new miniature chapter begins, with new feelings and thoughts. (Though I’m not sure how much other people’s thoughts depend on their surroundings, so maybe for others a change of context is less of a reset).
- Similarly, longer periods of repeatedly being in particular places with particular people. These might be decades of settled marriage or a few days of being on a trip. For me they are often a month to a year. They are punctuated by moving, breaking up, changing jobs. They tend to have their own routines and systems and patterns of thought. For me, starting a new one is often marked by a similar optimism and ambition for a fresh start as mornings. And ending one shares with evenings a risk of sadness at wasted opportunity.

Both of these also end because of something like sleep—changes of context that break the continuity of thoughts or habits within the period, either because those things relied on the previous context as something like memory, or because the new context asks for a new activity that replaces the old one, and the old one needed the continuity to stay alive.

[AN #75]: Solving Atari and Go with learned game models, and thoughts from a MIRI employee

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Highlights

Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model *(Julian Schrittwieser et al)* (summarized by Nicholas): Up until now, model-free RL approaches have been state of the art at visually rich domains such as Atari, while model-based RL has excelled for games which require planning many steps ahead, such as Go, chess, and shogi. This paper attains state of the art performance on Atari using a model-based approach, *MuZero*, while matching AlphaZero (AN #36) at Go, chess, and shogi while using less compute. Importantly, it does this without requiring any advance knowledge of the rules of the game.

MuZero's model has three components:

1. The *representation* function produces an initial internal state from all existing observations.

2. The *dynamics* function predicts the next internal state and immediate reward after taking an action in a given internal state.

3. The *prediction* function generates a policy and a value prediction from an internal state.

Although these are based on the structure of an MDP, **the internal states of the model do not necessarily have any human-interpretable meaning**. They are trained end-to-end only to accurately predict the policy, value function, and immediate reward. This model is then used to simulate trajectories for use in MCTS.

Nicholas's opinion: This is clearly a major step for model-based RL, becoming the state of the art on a very popular benchmark and enabling planning approaches to be used in domains with unknown rules or dynamics. I am typically optimistic about model-based approaches as progress towards safe AGI. They map well to how humans think about most complex tasks: we consider the likely outcomes of our actions and then plan accordingly. Additionally, model-based RL typically has the safety property that the programmers know what states the algorithm expects to pass through and end up in, which aids with interpretability and auditing. However, *MuZero* loses that property by using a learned model whose internal states are not constrained to have

any semantic meaning. I would be quite excited to see follow up work that enables us to understand what the model components are learning and how to audit them for particularly bad inaccuracies.

Rohin's opinion: *Note: This is more speculative than usual.* This approach seems really obvious and useful in hindsight (something I last felt for population-based training of hyperparameters). The main performance benefit (that I see) of model-based planning is that it only needs to use the environment interactions to learn how the environment *works*, rather than how to *act optimally* in the environment -- it can do the "act optimally" part using some MDP planning algorithm, or by simulating trajectories from the world model rather than requiring the actual environment. Intuitively, it should be significantly easier to learn how an environment works -- consider how easy it is for us to learn the rules of a game, as opposed to playing it well. However, most model-based approaches force the learned model to learn features that are useful for predicting the state, which may not be the ones that are useful for playing well, which can handicap their final performance. Model-free approaches on the other hand learn exactly the features that are needed for playing well -- but they have a much harder learning task, so it takes many more samples to learn, but can lead to better final performance. Ideally, we would like to get the benefits of using an MDP planning algorithm, while still only requiring the agent to learn features that are useful for acting optimally.

This is exactly what MuZero does, similarly to this previous paper: its "model" only predicts actions, rewards, and value functions, all of which are much more clearly relevant to acting optimally. However, the tasks that are learned from environment interactions are in some sense "easier" -- the model only needs to predict, *given a sequence of actions*, what the immediate reward will be. It notably *doesn't* need to do a great job of predicting how an action now will affect things ten turns from now, as long as it can predict how things ten turns from now will be *given* the ten actions used to get there. Of course, the model does need to predict the policy and the value function (both hard and dependent on the future), but the learning signal for this comes from MCTS, whereas model-free RL relies on credit assignment for this purpose. Since MCTS can consider multiple possible future scenarios, while credit assignment only gets to see the trajectory that was actually rolled out, we should expect that MCTS leads to significantly better gradients and faster learning.

I'm Buck Shlegeris, I do research and outreach at MIRI, AMA *(Buck Shlegeris)* (summarized by Rohin): Here are some beliefs that Buck reported that I think are particularly interesting (selected for relevance to AI safety):

1. He would probably not work on AI safety if he thought there was less than 30% chance of AGI within 50 years.

2. The ideas in Risks from Learned Optimization (AN #58) are extremely important.

3. If we build "business-as-usual ML", there will be inner alignment failures, which can't easily be fixed. In addition, the ML systems' goals may accidentally change as they self-improve, obviating any guarantees we had. The only way to solve this is to have a clearer picture of what we're doing when building these systems. *(This was a response to a question about the motivation for MIRI's research agenda, and so may not reflect his actual beliefs, but just his beliefs about MIRI's beliefs.)*

4. Different people who work on AI alignment have radically different pictures of what the development of AI will look like, what the alignment problem is, and what solutions

might look like.

5. Skilled and experienced AI safety researchers seem to have a much more holistic and much more concrete mindset: they consider a solution to be composed of many parts that solve subproblems that can be put together with different relative strengths, as opposed to searching for a single overall story for everything.

6. External criticism seems relatively unimportant in AI safety, where there isn't an established research community that has already figured out what kinds of arguments are most important.

Rohin's opinion: I strongly agree with 2 and 4, weakly agree with 1, 5, and 6, and disagree with 3.

Technical AI alignment

Problems

Defining AI wireheading *(Stuart Armstrong)* (summarized by Rohin): This post points out that "wireheading" is a fuzzy category. Consider a weather-controlling AI tasked with increasing atmospheric pressure, as measured by the world's barometers. If it made a tiny dome around each barometer and increased air pressure within the domes, we would call it wireheading. However, if we increase the size of the domes until it's a dome around the entire Earth, then it starts sounding like a perfectly reasonable way to optimize the reward function. Somewhere in the middle, it must have become unclear whether or not it was wireheading. The post suggests that wireheading can be defined as a subset of specification gaming (AN #1), where the "gaming" happens by focusing on some narrow measurement channel, and the fuzziness comes from what counts as a "narrow measurement channel".

Rohin's opinion: You may have noticed that this newsletter doesn't talk about wireheading very much; this is one of the reasons why. It seems like wireheading is a fuzzy subset of specification gaming, and is not particularly likely to be the only kind of specification gaming that could lead to catastrophe. I'd be surprised if we found some sort of solution where we'd say "this solves all of wireheading, but it doesn't solve specification gaming" -- there don't seem to be particular distinguishing features that would allow us to have a solution to wireheading but not specification gaming. There can of course be solutions to particular kinds of wireheading that *do* have clear distinguishing features, such as reward tampering (AN #71), but I don't usually expect these to be the major sources of AI risk.

Technical agendas and prioritization

The Value Definition Problem *(Sammy Martin)* (summarized by Rohin): This post considers the Value Definition Problem: what should we make our AI system try to do (AN #33) to have the best chance of a positive outcome? It argues that an answer to the problem should be judged based on how much easier it makes alignment, how competent the AI system has to be to optimize it, and how good the outcome would be if it was optimized. Solutions also differ on how "direct" they are -- on one end, explicitly writing down a utility function would be very direct, while on the other, something like Coherent Extrapolated Volition would be very indirect: it delegates the task of figuring out what is good to the AI system itself.

Rohin's opinion: I fall more on the side of preferring indirect approaches, though by that I mean that we should delegate to future humans, as opposed to defining some particular value-finding mechanism into an AI system that eventually produces a definition of values.

Miscellaneous (Alignment)

Self-Fulfilling Prophecies Aren't Always About Self-Awareness *(John Maxwell)* (summarized by Rohin): Could we prevent a superintelligent oracle from making self-fulfilling prophecies by preventing it from modeling itself? This post presents three scenarios in which self-fulfilling prophecies would still occur. For example, if instead of modeling itself, it models the fact that there's some AI system whose predictions frequently come true, it may try to predict what that AI system would say, and then say that. This would lead to self-fulfilling prophecies.

Analysing: Dangerous messages from future UFAI via Oracles and Breaking Oracles: hyperrationality and acausal trade *(Stuart Armstrong)* (summarized by Rohin): These posts point out a problem with counterfactual oracles (AN #59): a future misaligned agential AI system could commit to helping the oracle (e.g. by giving it maximal reward, or making its predictions come true) even in the event of an erasure, as long as the oracle makes predictions that cause humans to build the agential AI system. Alternatively, multiple oracles could acausally cooperate with each other to build an agential AI system that will reward all oracles.

AI strategy and policy

AI Alignment Podcast: Machine Ethics and AI Governance *(Lucas Perry and Wendell Wallach)* (summarized by Rohin): Machine ethics has aimed to figure out how to embed ethical reasoning in automated systems of today. In contrast, AI alignment starts from an assumption of intelligence, and then asks how to make the system behave well. Wendell expects that we will have to go through stages of development where we figure out how to embed moral reasoning in less intelligent systems before we can solve AI alignment.

Generally in governance, there's a problem that technologies are easy to regulate early on, but that's when we don't know what regulations would be good. Governance has become harder now, because it has become very crowded: there are more than 53 lists of principles for artificial intelligence and lots of proposed regulations and laws. One potential mitigation would be **governance coordinating committees**: a sort of issues manager that keeps track of a field, maps the issues and gaps, and figures out how they could be addressed.

In the intermediate term, the worry is that AI systems are giving increasing power to those who want to manipulate human behavior. In addition, job loss is a real issue. One possibility is that we could tax corporations relative to how many workers they laid off and how many jobs they created.

Thinking about AGI, governments should probably not be involved now (besides perhaps funding some of the research), since we have so little clarity on what the problem is and what needs to be done. We do need people monitoring risks, but there’s a pretty robust existing community doing this, so government doesn't need to be involved.

Rohin's opinion: I disagree with Wendell that current machine ethics will be necessary for AI alignment -- that might be the case, but it seems like things change significantly once our AI systems are smart enough to actually understand our moral systems, so that we no longer need to design special procedures to embed ethical reasoning in the AI system.

It does seem useful to have coordination on governance, along the lines of governance coordinating committees; it seems a lot better if there's only one or two groups that we need to convince of the importance of an issue, rather than 53 (!!).

Other progress in AI

Reinforcement learning

Learning to Predict Without Looking Ahead: World Models Without Forward Prediction *(C. Daniel Freeman et al)* (summarized by Sudhanshu): One critique of the World Models (AN #23) paper was that in any realistic setting, you only want to learn the features that are important for the task under consideration, while the VAE used in the paper would learn features for state reconstruction. This paper instead studies world models that are trained directly from reward, rather than by supervised learning on observed future states, which should lead to models that only focus on task-relevant features. Specifically, they use *observational dropout* on the environment percepts, where the true state is passed to the policy with a peek probability p, while a neural network, **M**, generates a proxy state with probability $1 - p$. At the next time-step, **M** takes the same input as the policy, plus the policy's action, and generates the next proxy state, which then may get passed to the controller, again with probability $1 - p$.

They investigate whether the emergent 'world model' **M** behaves like a good forward predictive model. They find that even with very low peek probability e.g. $p = 5\%$, **M** learns a good enough world model that enables the policy to perform reasonably well. Additionally, they find that world models thus learned can be used to train policies that sometimes transfer well to the real environment. They claim that the world model only learns features that are useful for task performance, but also note that interpretability of those features depends on inductive biases such as the network architecture.

Sudhanshu's opinion: This work warrants a visit for the easy-to-absorb animations and charts. On the other hand, they make a few innocent-sounding observations that made me uncomfortable because they weren't rigourously proved nor labelled as speculation, e.g. a) "At higher peek probabilities, the learned dynamics model is not needed to solve the task thus is never learned.", and b) "Here, the world model clearly only learns reliable transition maps for moving down and to the right, which is sufficient."

While this is a neat bit of work well presented, it is nevertheless still unlikely this (and most other current work in deep model-based RL) will scale to more complex alignment problems such as Embedded World-Models (AN #31); these world models do not capture the notion of an agent, and do not model the agent as an entity making long-horizon plans in the environment.

Deep learning

SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver *(Po-Wei Wang et al)* (summarized by Asya): Historically, deep learning architectures have struggled with problems that involve logical reasoning, since they often impose non-local constraints that gradient descent has a hard time learning. This paper presents a new technique, SATNet, which allows neural nets to solve logical reasoning problems by encoding them explicitly as MAXSAT-solving neural network layers. A MAXSAT problem provides a large set of logical constraints on an exponentially large set of options, and the goal is to find the option that satisfies as many logical constraints as possible. Since MaxSAT is NP-complete, the authors design a layer that solves a relaxation of the MaxSAT problem in its forward pass (that can be solved quickly, unlike MaxSAT), while the backward pass computes gradients as usual.

In experiment, SATNet is given bit representations of 9,000 9 x 9 Sudoku boards which it uses to learn the logical constraints of Sudoku, then presented with 1,000 test boards to solve. SATNet vastly outperforms traditional convolutional neural networks given the same training / test setup, achieving 98.3% test accuracy where the convolutional net achieves 0%. It performs similarly well on a "Visual" Sudoku problem where the trained network consists of initial layers that perform digit recognition followed by SATNet layers, achieving 63.2% accuracy where the convolutional net achieves 0.1%.

Asya's opinion: My impression is this is a big step forward in being able to embed logical reasoning in current deep learning techniques. From an engineering perspective, it seems extremely useful to be able to train systems that encorporate these layers end-to-end. It's worth being clear that in systems like these, a lot of generality is lost since part of the network is explicitly carved out for solving a particular problem of logical constraints-- it would be hard to use the same network to learn a different problem.

News

AI Safety Unconference 2019 *(David Krueger, Orpheus Lummis, and Gretchen Krueger)* (summarized by Rohin): Like last year, there will be an AI safety unconference alongside NeurIPS, on Monday Dec 9 from 10am to 6pm. While the website suggests a registration deadline of Nov 25, the organizers have told me it's a soft deadline, but you probably should register now to secure a place.

Neural Annealing: Toward a Neural Theory of Everything (crosspost)

The following is QRI's unified theory of music, meditation, psychedelics, depression, trauma, and emotional processing. Implications for how the brain implements Bayesian updating, and future directions for neuroscience. Crossposted from http://opentheory.net

---

Context: follow-up to The Neuroscience of Meditation and A Future For Neuroscience; a unification of (1) *the Entropic Brain* & *REBUS* (Carhart-Harris et al. 2014; 2018; 2019), (2) *the Free Energy Principle* (Friston 2010), (3) *Connectome-Specific Harmonic Waves* (Atasoy et al. 2016; 2017), and (4) QRI's *Symmetry Theory of Valence* (Johnson 2016; Gomez Emilsson 2017).

0. Introduction

Why is neuroscience so hard?

Part of the problem is that the brain is complicated. But we've also mostly been doing it wrong, trying to explain the brain using methods that couldn't possibly generate insight about the things we care about.

On QRI's lineages page, we suggest there's a distinction between 'old' and 'new' neuroscience:

> *Traditionally, neuroscience has been concerned with cataloguing the brain, e.g. collecting discrete observations about anatomy, observed cyclic patterns (EEG frequencies), and cell types and neurotransmitters, and trying to match these facts with functional stories. However, it's increasingly clear that these sorts of neat stories about localized function are artifacts of the tools we're using to look at the brain, not of the brain's underlying computational structure.*
>
> *What's the alternative? Instead of centering our exploration on the sorts of raw data our tools are able to gather, we can approach the brain as a self-organizing system, something which uses a few core principles to both build and regulate itself. As such, if we can reverse-engineer these core principles and use what tools we have to validate these bottom-up models, we can both understand the internal logic of the brain's algorithms — the how and why the brain does what it does — as well as find more elegant intervention points for altering it.*

That's a big check to try to cash. What might this look like?

I. Annealing metaphors for the brain

In my post about the neuroscience of meditation, I talked about *simulated annealing*, a natural implication of Robin Carhart-Harris's work on entropic disintegration in the brain:

> *Annealing involves heating a metal above its recrystallization temperature, keeping it there for long enough for the microstructure of the metal to reach*

equilibrium, then slowly cooling it down, letting new patterns crystallize. This releases the internal stresses of the material, and is often used to restore ductility (plasticity and toughness) on metals that have been 'cold-worked' and have become very hard and brittle— in a sense, annealing is a 'reset switch' which allows metals to go back to a more pristine, natural state after being bent or stressed. I suspect this is a useful metaphor for brains, in that they can become hard and brittle over time with a build-up of internal stresses, and these stresses can be released by periodically entering high-energy states where a more natural neural microstructure can reemerge.

In his work on the entropic brain, Carhart-Harris studies how psychedelics like LSD and psilocybin add enough energy (neural activity) to the brain that existing neural patterns are disrupted, much like how heating a metal disrupts its existing molecular bonds. Recently, Carhart-Harris and Friston have unified their frameworks under the REBUS (RElaxed Beliefs Under pSychedelics) model, which also imports the annealing metaphor for brains:

The hypothesized flattening of the brain's (variational free) energy landscape under psychedelics can be seen as analogous to the phenomenon of simulated annealing in computer science—which itself is analogous to annealing in metallurgy, whereby a system is heated (i.e., instantiated by increased neural excitability), such that it attains a state of heightened plasticity, in which the discovery of new energy minima (relatively stable places/trajectories for the system to visit/reside in for a period of time) is accelerated (Wang and Smith, 1998). Subsequently, as the drug is metabolized and the system cools, its dynamics begin to stabilize—and attractor basins begin to steepen again (Carhart-Harris et al., 2017). This process may result in the emergence of a new energy landscape with revised properties.

It's a powerful metaphor since it ties together and recontextualizes so many core neuroscience concepts: free energy landscapes, Bayesian modeling, the 'handshake' between bottom-up sense-data and top-down priors. For a general overview of the math, see Wikipedia on simulated annealing, Metropolis-Hastings algorithm, Parallel tempering; for more on Carhart-Harris's and Friston's work, see Scott Alexander's and Milan Griffes' commentary. There seems to be some convergence on this metaphor: as Scott Alexander noted,

F&CH aren't the first people to discuss this theory of psychedelics. It's been in the air for a couple of years now – and props to local bloggers at the Qualia Research Institute and Mad.Science.Blog for getting good explanations up before the parts had even all come together in journal articles. I'm especially interested in QRI's theory that meditation has the same kind of annealing effect, which I think would explain a lot.

The basics: how does annealing work?

Carhart-Harris's and Friston's model does many very clever things and is a substantial addition to the literature; I start from a similar frame but describe the process slightly differently. The following is QRI's model (based on my talk on the Neuroscience of Meditation in Thailand):

- First, energy (neural excitation, e.g. Free Energy from prediction errors) builds up in the brain, either gradually or suddenly, collecting disproportionately in the brain's natural eigenmodes;

- This build-up of energy (rate of neural firing) crosses a metastability threshold and the brain enters a high-energy state, causing entropic disintegration (weakening previously 'sticky' attractors);
- The brain's neurons self-organize into new multi-scale equilibria (attractors), aka implicit assumptions about reality's structure and value weightings, which given present information should generate lower levels of prediction error than previous models (this is implicitly both a *resynchronization of internal predictive models with the environment*, and a *minimization of dissonance in connectome-specific harmonic waves*);
- The brain 'cools' (neural activity levels slowly return to normal), and parts of the new self-organized patterns remain and become part of the brain's normal activity landscape;
- The cycle repeats, as the brain's models become outdated and prediction errors start to build up again.

Any 'emotionally intense' experience that you need time to process most likely involves this *entropic disintegration->search->annealing* mechanism— this is what emotional processing *is*.

And I'd suggest that this is the *core dynamic of how the brain updates its structure*, the mechanism the brain uses to pay down its 'technical debt'. In other words, entering high-energy states (i.e., intense emotional states which take some time to 'process') is how the brain releases structural stress and adapts to new developments. This process needs to happen on a regular basis to support healthy function, and if it doesn't, psychological health degrades— In particular, mental flexibility & emotional vibrancy go down — analogous to a drop in a metal's 'ductility'. People seem to have a strong subconscious drive toward entering these states and if they haven't experienced a high-energy brain state in some time, they actively seek one out, even sometimes in destructive ways.

However, the brain spends most of its time in low-energy states, because they're safer: systems in noisy environments need to limit their rate of updating. There are often *spikes* of energy in the brain, but these don't tend to snowball into full high-energy states because the brain has many 'energy sinks' (inhibitory top-down predictive models) which soak up excess energy before entropic disintegration can occur.

But the brain can enter high-energy states if these energy sinks are:

(1) <u>De-activated</u>, if certain evolved trigger conditions are present- e.g., death of a loved one, falling in love, good sex, social rejection, getting bitten by a weird animal, failing some important prediction. In these cases there seems to be some sort of adaptive gating mechanism that disables the typical energy sinks in order to allow entropic disintegration->search->annealing to happen.

(2) <u>Overwhelmed</u>, if there's an enormous magnitude of energy coming in, faster than the energy sinks can mop it up- e.g., watching a horror movie, direct brain stimulation, first day of school, being sleep deprived, military boot camp, cult indoctrinations, your wedding day.

(3) <u>Avoided</u>, if *semantically-neutral energy* is applied to the system. Essentially, coherent energy which isn't strongly linked to any cognitive, emotional, or sensory process will be partially illegible to most existing energy sinks, and so it can persist

long enough to build up – basically 'hacking' the brain's activity normalization system. (Hold that thought – this is the most interesting one. We'll return to it later.)

This is the 'view from 30,000 feet' for how simulated annealing in the brain works. If you stopped reading here, you'd walk away with a reasonable toy model of QRI's "Neural Annealing" framework.

But there's a lot more to the model! The rest of this writeup is an iterative tour using Neural Annealing to explain meditation, trauma, love, depression, psychedelics, and effective therapy, with each section adding a variation on the core theme.

Interlude: FEP, CSHW, and EBH/REBUS

QRI's "Neural Annealing" framework is essentially a unification of Karl Friston's Free Energy Principle (FEP), Selen Atasoy's Connectome-Specific Harmonic Waves (CSHW), Robin Carhart-Harris's Entropic Brain Hypothesis (EBH), and QRI's own Symmetry Theory of Valence (STV). Recently, Friston and Carhart-Harris have unified their respective paradigms with the *Relaxed Beliefs Under pSychedelics* (REBUS) model. I believe combining *all three* is exponentially more powerful, not only giving the *computational-level story* of REBUS, but also giving us a model for *how the brain may be physically implementing REBUS, and Bayesian updating in general*, with a correspondingly richer set of predictions.

First, here's a quick recap: to paraphrase what I wrote elsewhere,

Karl Friston's Free Energy Principle (FEP) is the leading theory of self-organizing system dynamics, one which has (in various guises) pretty much taken neuroscience by storm. It argues that any self-organizing system which effectively resists disorder must (as its core organizing principle) minimize its free energy, that free energy is equivalent to surprise (in a Bayesian sense), and that this surprise-minimization drives basically all human behavior. This minimization of surprise revolves around Bayesian-type reasoning: the brain is always getting bottom-up sense data flowing in, more than it can handle. So it relies on top-down predictive models that attempt to sort through all this data so we can focus on the surprising stuff, the stuff that can't be effortlessly predicted. The core of the FEP is the details of how this 'handshake' between bottom-up and top-down happens, and what can influence it. See Friston's primary work; Scott Alexander's attempt to distill it. Related to (and sometimes used synonymously with) Active Inference, the Bayesian Brain, and Predictive Processing / Predictive Coding.

Robin Carhart-Harris's Entropic Brain Hypothesis (EBH) is essentially an attempt to import key concepts such as entropy and self-organized criticality from statistical physics into neuroscience, in order to explain psychedelic phenomena. As I noted above, it suggests that certain conditions *such as* psychedelics can add enough energy to brain networks that they undergo 'entropic disintegration', and then self-organize into new equilibria. See Carhart-Harris 2018.

Selen Atasoy's Connectome-Specific Harmonic Waves (CSHW) is a method for applying harmonic analysis to the brain: basically, it uses various forms of brain imaging to infer what the brain's natural resonant frequencies (eigenmodes) are, and how much energy each of these frequencies have. The core workflow is three steps: first combine MRI and DTI to approximate a brain's connectome, then with an empirically-derived wave propagation equation calculate what the natural harmonics are of this connectome, then estimate which power distribution between these harmonics would most accurately reconstruct the observed fMRI activity. This

framework offers several notable things: (a) these connectome-specific harmonic waves (CSHWs) are natural Schelling points that the brain has probably self-organized around (and so are worth talking about); (b) a plausible mid-level bridge connecting bottom-up neural dynamics and high-level psychological phenomena, (c) something we can actually measure. CSHW is an empirical paradigm, which is very uncommon in theoretical neuroscience. Here's a transcript of Atasoy's explanation; I also wrote extensively about CSHW in A Future for Neuroscience.

In short: each of these three paradigms is a description of how the brain self-organizes. Friston's work understands the self-organization from a computational lens; Carhart-Harris an energetic lens; Atasoy a physical lens.

Finally, I'd offer two further pieces of background context:

QRI's own Symmetry Theory of Valence (STV), which hypothesizes that given a mathematical representation of an experience, the *symmetry* of this representation will encode how pleasant the experience is (Johnson 2016). We further hypothesize that consonance between a brain's connectome-specific harmonic waves (CSHWs) will be a reasonable proxy for this symmetry (Gomez Emilsson 2017).

Marr's Three Levels: as explained on our lineages page,

> ***David Marr*** *is most famous for Marr's Three Levels (along with Tomaso Poggio), which describe "the three levels at which any machine carrying out an information-processing task must be understood:"*
> *>Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?*
> *>Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?*
> *>Hardware implementation: How can the representation and algorithm be realized physically? [Marr (1982), p. 25]*
> *This framework sounds simple, but is remarkably important since arguably most of the confusion in neuroscience (and phenomenology research) comes from starting a sentence on one Marr-Poggio level and finishing it on another, and this framework lets people debug that confusion.*

Back to annealing –

As noted, Carhart-Harris and Friston have unified their paradigms under REBUS by understanding prediction errors as the 'energy' parameter which drives disruption (entropic disintegration) in the brain's networks. Over time, this drives an evolutionary search function which attempts to minimize these prediction errors. I think this is a very beautiful description of a very clever system, and one which allows us an opportunity to cross-validate each model, and jump between levels of description if we get 'stuck'. But it's still missing a story about physical implementation. *What is this 'energy', physically speaking?*

II. How meditation works: semantically-neutral annealing

I believe that almost all techniques that intentionally 'hack' the brain's annealing process share a common mechanism: a build-up of *semantically neutral energy*. "Semantically neutral energy" refers to neural activity which is not strongly associated with any specific cognitive or emotional process. As I note above, usually energy build-up is limited: once a perturbation of the system neatly falls into a pattern

recognized by the brain's predictive hierarchy, the neural activity propagating this pattern is dissipated. But if a pattern *never quite matches* anything, or takes advance of implementation-level structure to persist, and especially if it's getting continually reinforced by some external or internal dynamic it can persist long enough to build up. I think meditation is a perfect example of a process which adds semantically-neutral energy to the brain: effortful attention on excitatory bottom-up sense-data and attenuation of inhibitory top-down predictive models will naturally lead to a build-up of this 'non-semantic' energy in the brain. From The Neuroscience of Meditation:

> *Furthermore, from what I gather from experienced meditators, successfully entering meditative flow may be one of the most reliable ways to reach these high-energy brain states. I.e., it's very common for meditation to produce feelings of high intensity, at least in people able to actually enter meditative flow. Meditation also produces more 'pure' or 'neutral' high-energy states, ones that are free of the intentional content usually associated with intense experiences which may distort or limit the scope of the annealing process. So we can think of intermediate-to-advanced ('successful flow-state') meditation as a reheating process, whereby the brain enters a more plastic and neutral state, releases pent-up structural stresses, and recrystallizes into a more balanced, neutral configuration as it cools. Iterated many times, this will drive an evolutionary process and will produce a very different brain, one which is more unified & anti-fragile, less distorted toward intentionality, and in general structurally optimized against stress.*
>
> *An open question is how or why meditation produces high-energy brain states. There isn't any consensus on this, but with a nod to the predictive coding framework, I'd offer that bottom-up sense-data is generally excitatory, adding energy to the system, whereas top-down predictive Bayesian models are generally inhibitory, functioning as 'energy sinks'. And so by 'noting and knowing' our sensations before our top-down models activate, in a sense we're diverting the 'energy' of our sensations away from its usual counterbalancing force. If we do this long enough and skillfully enough, this energy can build up and lead to 'entropic disintegration', essentially pushing enough energy into the system that existing attractors are disrupted and annealing can occur.*

A natural question here is *what *is* this 'semantically neutral energy' exactly?* – an abstract answer here is "semantically neutral energy" can be thought of as an increase in brain activity which is (1) illegible to Marr's semantic/computational level, but (2) coherent with regard to Marr's algorithmic or implementational levels (another term for this might be 'semantically-illegible energy'). But my concrete answer is that *semantically neutral energy is a build-up of energy in the brain's natural resonances* — energy accumulating in CSHWs. And so it's *this* that builds up during meditation, and *this* that starts a *semantically-neutral annealing* process which has a unique effect profile.

I think *semantically-neutral annealing* is the *best kind* of annealing for psychological health, because:

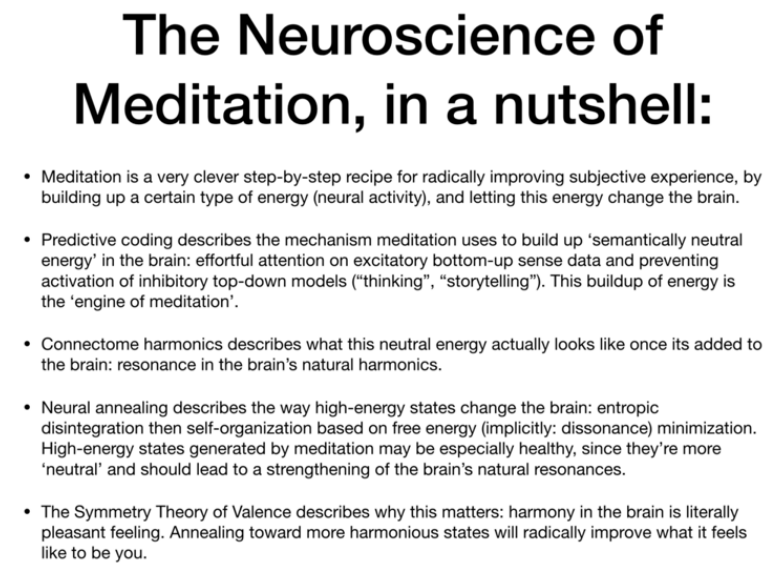
(1) By mostly avoiding energy sinks, the same entropic disintegration->search->annealing process can happen using less total energy, which is less disruptive to the fine details of the system;

(2) Since this energy is semantically-neutral, it doesn't depend on or trigger as many semantic processes in the brain (which can have unpredictable effects), and likewise it

doesn't necessarily rely on anti-inductive 'hacks' to trick the predictive processing system, and these factors make it a more reliable and repeatable source of annealing;

(3) *Very very importantly:* similarly to how vibratory energy applied to a tuning fork quickly collapses to the natural resonant frequency of the tuning fork, I'm speculating that coherent, semantically-neutral energy added to the brain will naturally cluster in the brain's natural connectome harmonics, which will thus drive an annealing process which strengthens a consonant subset of the brain's natural harmonic resonances in the long-term— essentially 'retuning the brain' toward more resonant/flow states. For more details, see The Neuroscience of Meditation;

(4) Finally, this process should *feel really really good* and in the long-term, retune the mind to be more pleasant to inhabit. QRI's work on the Symmetry Theory of Valence (STV) and our method of applying this to the brain (CDNS) suggests that harmony in the brain is *literally synonymous* with pleasure, and so processes which 'deepen the grooves' of core harmonic resonances will tend to boost the mind's default hedonic level (likely helping significantly with neuroticism and emotional resilience).



I.e., Meditation is a remarkably clever technique which piggybacks on several of the brain's core principles of self-organization: first, effortful attention on (excitatory) sense-data and inhibiting (inhibitory) predictive storytelling naturally pushes the brain into a high-energy state and makes it more malleable; this excess energy disproportionately collects in natural brain harmonics, and as the brain 'cools' from its high-energy state, these energized harmonics become 'deeper', leading to more psychological robustness. Less neuroticism and more flow. *I think this is where a large portion of the benefits of advanced meditation comes from.*

Meditation isn't the only method to induce build-up of semantically-neutral energy; the "Big Three" are:

– **Meditation**, which seems to work by both increasing excitatory sense-data and decreasing inhibitory top-down predictive models (energy sinks);

– **Psychedelics**, which intuitively may function by disabling existing energy sinks (or perhaps overloading them by increasing baseline firing rates or increasing the branching factor of neural activity).

– **Music**, a sensory input which seems to exist on the knife's edge between exhibiting highly ordered patterns (some of which will hit natural connectome harmonics and so allow accumulation of energy through resonance) on one hand, and on the other hand not being *too* predictable (thus dodging most inhibitory top-down predictive models);

Hybrid approaches also exist: e.g. exercise, dance, sex, tantric practices, EMDR, and breath work are essentially combinations of the rhythmic portion of music and the sensory portion of meditation. The fact that psychedelics reliably enhance the potency of each and every one of these practices is not a coincidence, but due to shared mechanism.[1]

III. Depression as a disorder of annealing; bipolar depression doubly so

To describe depression in one sentence: *"Depression is a self-reinforcing perturbation from the natural annealing cycle."* There are two related aspects to this: (1) an inability to anneal normally, and (2) annealing abnormally (more specifically, annealing new attractor basins which are high in dissonance, or annealing a pathological change in energy parameter dynamics).

Most people have a simple model of depression as "being sad all the time" – but I think a two-factor model looking at *energy parameter* and *valence* offers a lot of clarity and predictive utility. Roughly speaking, this suggests parametrizing depression into three core types:

I. Depression with no high energy states, characterized by a lack of annealing (emotional clarity and dynamism) in general;

II. Depression with high-energy negative states, which over time anneals minds toward *suffering* and *hopelessness*;

III. Bipolar depression with high-energy positive & negative states, which over time anneals minds toward the *dramatic*.

These categories aren't exclusive or static; too much time in one will increase the probability one may also fall into the others.

Not annealing frequently enough may be the most important 'non-obvious' cause of depression. Brains – especially younger ones, since they're changing so much – *really do need to anneal regularly* to pay down their 'technical debt', and if they don't, they grow brittle and neurotic. (Technical debt in the brain builds up as we twist our existing brain networks to accommodate new facts; this debt is 'paid down' when we enter high-energy states and let new brain networks which fit these constraints self-organize) The 'annealing pressure' also increases over time, and if a wholesome annealing opportunity fails to present itself, the brain will progressively lower its standards looking for *any* opportunity for annealing. Especially if done repeatedly, this can cause long-term damage to the brain's attractor basin landscape. (We see this in negative coping strategies such as cutting, drama-seeking, and so on – if someone is engaging in such, they've probably annealed poorly, and also likely have few realistic opportunities for healthy annealing.) Many forms of entertainment we think of as palliatives in today's society (e.g. movies, video games) may be weak-and-

incomplete-but-still-nonzero drivers of annealing. Not as good as the real thing, but better than nothing if that's your only option.

At the high-energy extreme, it seems likely and tragic that depression compounds itself by repeatedly causing intense negative emotion (high-energy states) which anneals the brain toward these patterns, and toward assigning salience on the set of problems and types of thoughts (attractor landscape) facing a depressed person — many of which are their own cause and would weaken if ignored. Relatedly, I suspect some CSHW- and music-theory-related math could be found describing how depression anneals what I would call a brain's 'connectome key signature' (CKS) toward a 'minor key', an internal logic which feels tragic/hopeless (has fewer harmonious arrangements and progressions), which the brain then uses as building blocks for its reality.

Bipolar depression seems a little more strange; the extreme highs and lows may in aggregate produce crazier annealing patterns than just one or the other — essentially there's a 'tug of war' between patterns annealed during each extreme, which prioritizes the survival of the class of patterns that exist during *both* extremely positive and extremely negative states. In practice, over time this anneals a mind's stories toward the dramatic, and toward reducing the activation energy needed to flip the brain between major and minor keys (the psych literature calls this 'kindling'). Each of these 'key signature flips' would itself release a great deal of pent-up energy, further driving the annealing process. As I note in A Future for Neuroscience:

> *This is not to say our key signatures are completely static, however: an interesting thread to pull here may be that some brains seem to flip between a major key and a minor key, with these keys being local maximas of harmony. I suspect each is better at certain kinds of processing, and although parts of each can be compatible with the other, each has elements that present as defection to the internal logic of the other and so these attractors can be 'sticky'. But there can also be a buildup of tension as one gathers information that is incompatible with one's key signature, which gets progressively more difficult to maintain, and can lead to the sort of intensity of experience that drives an annealing-like process when the key signature flips. And in the case of repeated flips, the patterns which are compatible with both key signatures will be the most strongly reinforced.*

In some ways a bipolar brain may result in significant cognitive and creative advantages: perhaps the biggest is more access to high-energy states, which in the short term helps creativity by allowing more exploration and also steeper valence gradients to follow, and iterated over the long term allows significantly more optimization pressure on the subsystems that are repeatedly annealed. However this has corresponding epistemological downsides as noted above; fueling creative work with valence deltas is likely to 'warp the engine' over time, to paraphrase Shinzen. Friston's notion that 'systems maximizing long-term stability spend most of their time in a small number of states' seems particularly relevant to mood disorders. (My colleague Andrés suggests this 'bipolar effect profile' may be replicated by valence-enhancing drugs with a short duration and hangover, such as cocaine- this at least fits stereotypes.)

I find myself wondering if neuroticism can be thought of as ancient neural technology intended to *reduce annealing frequency* in the ancestral environment — essentially if we look into the brains of highly neurotic people, we might find strong energy sinks located around natural connectome harmonics which prevent semantically-neutral energy build-up. This likely contributes to certain forms of depression (and leads to

pernicious feedback cycles — the less one anneals, the more neurotic one gets, the less able to reach high-energy states one becomes), but might also help prevent seizures or inappropriate updating/annealing, and may have frequency-dependent benefits. E.g., a group with 19 carefree annealers and 1 neurotic guardian will act more wisely than one with 20 carefree annealers or 20 neurotic guardians. The 'neuroticism=energy sinks' frame seems to suggest how to reduce neuroticism (anneal more often, especially semantically-neutral annealing), and also offer clues as to how neuroticism is implemented in the brain: we might look into the mathematics of Anderson localization in the connectome: topological features that can 'eat' waves.

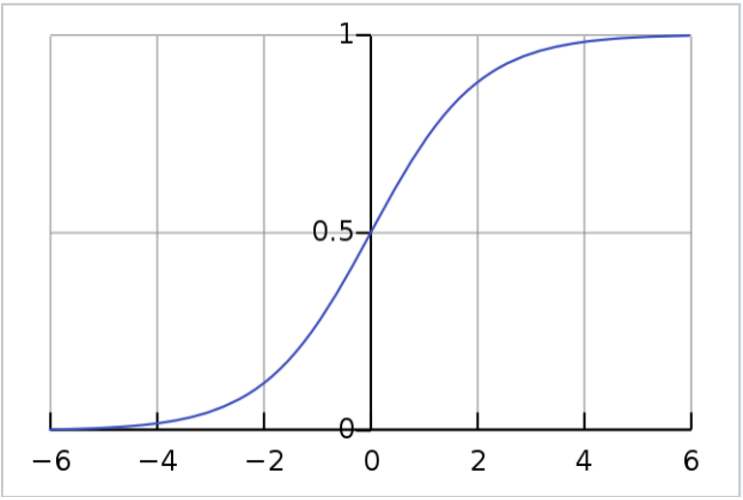
Is sleep a natural annealing process? If so, this could cleanly explain the connection between depression and chronic sleep disturbances — poor sleep as both a cause and effect of infrequent annealing. And it would indicate a treatment path: a restoration of normal annealing patterns may help improve both mood and sleep. I hold the following lightly, but we might model nREM as the heating-up phase (undampened harmonics) and REM as the neural search & cooling process. From a review drawing parallels between sleep and jhana (intense meditative) states:

> *This paper is a preliminary report on the first detailed EEG study of jhana meditation, with findings radically different to studies of more familiar, less focused forms of meditation. While remaining highly alert and "present" in their subjective experience, a high proportion of subjects display "spindle" activity in their EEG, superficially similar to sleep spindles of stage 2 nREM sleep, while more-experienced subjects display high voltage slow-waves reminiscent, but significantly different, to the slow waves of deeper stage 4 nREM sleep, or even high-voltage delta coma. Some others show brief posterior spike-wave bursts, again similar, but with significant differences, to absence epilepsy. Some subjects also develop the ability to consciously evoke clonic seizure-like activity at will, under full control. (Dennison 2019)*

It seems plausible that broad rhythmic brain activity helps with certain 'physical housekeeping' tasks in the brain as well, and if one anneals regularly they may need somewhat less sleep (see recent research on Alzheimer's, sleep, and rhythmic stimulation helping break up brain plaques).

The 'dead neuron' model of neuroticism and depression:

Deep learning models can exhibit 'dead neurons': neurons whose activation function gets 'stuck' on the on or off position, for instance when a sigmoid function gets too high or too low and its slope drops to almost zero. These 'dead' neurons can be nigh-impossible to 'revive' within the model, since it can be the case that their gradient (implicit sensitivity to input) is so shallow that there simply aren't inputs that will nudge it in one or another direction.



Graphic: sigmoidal function. This loses sensitivity when values get too high or too low. Different activation functions can lose sensitivity (lead to 'dead neurons') under different scenarios- ReLU is notorious for this.

These 'dead' neurons tend to cause lots of problems, since their "always-on" or "always off" signal tends to propagate through the network very strongly, causing later neurons in the chain to also exhibit less sensitivity to input. (Sometimes this process will cascade, sometimes not, much like malignant vs benign tumors.)

I suspect this might be a strong frame for understanding the 'psychological cruft' which builds up in brains, and how and why regular annealing is so healthy: over time, sensitive neurons can slide into this broken state, shifting from conditional values to the neurological equivalent of static 0s and 1s. In this case I would expect more neuroticism, less flexible thinking, lower emotional resilience, and worse epistemology from people who haven't annealed recently: lots of all-or-nothing thinking. But by injecting lots of energy into the system, enough of the internal and external context of these neurons is shifted such that some of them may get 'reset' and regain their conditional processing state. At the very least, this self-reorganization process can allow these neurons to move to less-critical points in processing networks.

An idea related to this frame is that a core function of neural annealing is to maintain a smooth gradient of harmony in the brain (and mind) – to make it possible to "follow your joy" toward better outcomes. If this breaks down and you can't "follow your joy", consider putting yourself in a situation which could plausibly kickstart an annealing process (even if you don't feel emotionally motivated to do so).

IV. The nature of trauma and the implementation of the Bayesian Brain

Trauma is one of the worst elements of the human condition. It's easy enough to accumulate that we all have some, and it's hard to get rid of. But *what is it?*

Scott Alexander recently reviewed a core work in the PTSD literature, *The Body Keeps The Score*, and offers some context:

> *The book stressed the variety of responses to PTSD. Some people get anxious. Some people get angry. But a lot of people, whatever their other symptoms, also go completely numb. They are probably still "having" "emotions" "under" "the" "surface", but they have no perception of them. Sometimes this mental deficit is*

accompanied by equally surprising bodily deficits. Van der Kolk describes a study on stereoagnosia in PTSD patients: if blindfolded and given a small object (like a key), they are unable to recognize it by feel, even though this task is easy for healthy people. Sometimes this gets even more extreme, like the case of a massage therapy patient who did not realize they were being massaged until the therapist verbally acknowledged she had started.

The book is called The Body Keeps The Score, and it returns again and again to the idea of PTSD patients as disconnected from their bodies. The body sends a rich flow of information to the brain, which is part of what we mean when we say we "feel alive" or "feel like I'm in my body". In PTSD, this flow gets interrupted. People feel "like nothing". …

There's some discussion of the neurobiology of all this, but it never really connects with the vividness of the anecdotes. A lot of stuff about how trauma causes the lizard brain to inappropriately activate in ways the rational brain can't control, how your "smoke detector" can be set to overdrive, all backed up with the proper set of big words like "dorsolateral prefrontal cortex" – but none of it seemed to reach the point where I felt like I was making progress to a gears-level explanation. I felt like the level on which I wanted an explanation of PTSD, and the level at which van der Kolk was explaining PTSD, never really connected; I can't put it any better than that. …

There are a lot of alternative treatments for PTSD. Neurofeedback, where you attach yourself to a machine that reads your brain waves and try to explore the effect your thoughts have on brain wave production until you are consciously able to manipulate your neural states. Internal family systems, where a therapist guides you through discovering "parts" of yourself (think a weak version of multiple personalities), and you talk to them, and figure out what they want, and make bargains with them where they get what they want and so stop causing mental illness. Eye movement directed reprocessing (alternative when the book was written, now basically establishment) where you move your eyes back and forth while talking about your trauma, and this seems to somehow help you process it better. Acupuncture. Massage. Yoga. …

Maybe the most consistent lesson from this book's tour of successful alternative therapies – keeping with the theme of the title – is that it's important for PTSD patients to get back in touch with their bodies. Massage therapy, yoga, and acupuncture addressed this directly, usually creating gentle, comfortable sensations that patients could take note of to gradually relax the absolute firewall between bodily sensation and conscious processing.

The simple Neural Annealing take on trauma is that significant negative events can push the brain into a high-energy state filled with 'trauma patterns', and as the brain cools, some of these trauma patterns crystallize/anneal in a very durable form, which present as PTSD.

I think this is a *more useful* answer than what's out there currently, offering straightforward intuitive answers for (1) what kinds of things are most likely to cause PTSD, (2) why PTSD is so 'sticky', and (3) an intuitive solution to PTSD: anneal over the bad patterns with better patterns.

But Scott's description seems to point at something further: that there's a disconnection happening with trauma. To address this, I propose the Neural Annealing

model for how CSHW could *implement* the Bayesian Brain model of cognition. We'll then circle back and discuss what might be going wrong during trauma.

Last year in A Future for Neuroscience, I shared the frame that we could split CSHWs into high-frequency and low-frequency types, and perhaps say something about how they might serve different purposes in the Bayesian brain:

> ***The mathematics of signal propagation and the nature of emotions***
> *High frequency harmonics will tend to stop at the boundaries of brain regions, and thus will be used more for fine-grained and very local information processing; low frequency harmonics will tend to travel longer distances, much as low frequency sounds travel better through walls. This paints a possible, and I think useful, picture of what emotions fundamentally are: semi-discrete conditional bundles of low(ish) frequency brain harmonics that essentially act as Bayesian priors for our limbic system. Change the harmonics, change the priors and thus the behavior. Panksepp's seven core drives (play, panic/grief, fear, rage, seeking, lust, care) might be a decent first-pass approximation for the attractors in this system.*

I would now add this roughly implies a *continuum* of CSHWs, with scale-free functional roles:

- Region-specific harmonic waves (**RSHWs**) – high frequency resonances that implement the processing of cognitive particulars, and are localized to a specific brain region (much like how high-frequencies don't travel through walls) – in theory quantifiable through simply applying Atasoy's CSHW method to individual brain regions;
- Connectome-specific harmonic waves (**CSHWs**) – low-frequency connectome-wide resonances that act as Bayesian priors, carrying relatively simple 'emotional-type' information across the brain;
- Sensorium-specific harmonic waves (**SSHWs**) – very-low-frequency waves that span not just the connectome, but the larger nervous system and parts of the body. These encode *somatic* information – in theory, we could infer sensorium eigenmodes by applying Atasoy's method to not only the connectome, but the nervous system, adjusting for variable nerve-lengths, and validate against something like body-emotion maps.[2][3]

These waves shade into each other – a 'low-frequency thought' shades into a 'high-frequency emotion', a 'low-frequency emotion' shades into somatic information. As we go further up in frequencies, these waves become more localized.

An interesting implication here is *we may essentially get Bayesian updating to naturally emerge from this typology*, through interactions between these various waves: essentially, I think it's 'injection-locking all the way down'. (Injection-locking is when harmonic oscillators (like CSHWs) essentially 'sync up' their periods and phases.) Specifically:

Low-frequency CSHWs carry priors, higher frequency RSHWs deal with particulars. Lower frequencies span the brain; higher frequencies resonate within more local regions of the brain — the higher the frequency of the wave, the smaller the region it tends to resonate in. The RSHWs in different regions can't talk to each other directly, since (definitionally) these waves can't travel across regional boundaries. But they *can* talk to each other indirectly, through interacting with low-frequency CSHWs. More specifically, I speculate that regions and CSHW-encoded priors interact through a **power-weighted averaging between CSHWs and RSHWs**, as mediated by the

math of injection-locking and injection-pulling. This **allows both functional partitioning and also global updating:** regions get some isolation in order to perform their specialized computations, but they also get exposure to data about the overall Bayesian prior situation, aka what we call 'emotional information'. I.e. Region A syncs up with CSHWs, which carry the information to Region B and sync up with the RSHWs there, and so on. Of note, there's a delicate, power-weighted handshake between CSHWs and RSHWs: low-frequency harmonics (emotions / Bayesian priors) carry more power per harmonic (lower due to frequency, much higher due to amplitude) but there are many more high-frequency harmonics (sensory+cognitive particulars). Strong emotions like anger likely pump huge amounts of energy into CSHWs and upend this balance, forcing RSHWs to synchronize with CSHWs. We can think of this as sacrificing the delicate epistemology-harmonization handshake in favor of unity of processing and clarity of action — or put simply, forcing perception to match top-down expectations.

On entropic disintegration, search, and annealing in evolved harmonic systems:

The noisy, stochastic nature of brain activity, along with practical requirements for homeostasis, will lead to a strong optimization of the CSHW+RSHW configuration for local minima which are resistant to change. However, a large enough perturbation will push the system out of this basin (**entropic disintegration** step). The **neural search** step is essentially the system stochastically testing different harmonic configurations; the **neural annealing** step is the system 'settling into' a configuration as its top-down predictive models get sufficiently good at sopping up the excess energy in the system, essentially forming a new basin it will again take a large amount of perturbation to get out of. The strength of annealing can be thought of as the steepness of this basin, and also the Hebbian reinforcement of system attractors ("neurons that fire together, wire together"). Insofar as partitioning is possible in a broadly-coupled harmonic system, these perturbations will tend to be 'local' as the brain has strong incentives to preserve structure that doesn't need updating.

Toward a generalized definition of trauma: a breakdown of information-propagation-via-injection-locking

I propose that sometimes the brain needs to rapidly halt information propagation across regions to prevent cascading system failure (a metaphor that comes to mind is an uncontrolled prion-like change in the local key signature that ripples out from a traumatized region, progressively breaking cybernetic calibrations). I believe the brain uses two interlinked mechanisms to do this: (1) weakening CSHWs, thus weakening information propagation throughout the brain, and (2) arranging different brain regions into frequency regimes which make information transfer difficult between them (the golden mean is the mathematically-optimal ratio for non-interaction). Once this happens, it can be very hard to reverse, since it forms a self-sustaining cycle: (1) causes (2) and (2) causes (1). *We call this 'trauma'.*

Some predictions from this – I'd expect to see substantially less energy in low-frequency CSHWs after trauma, and substantially more energy in low-frequency CSHWs during both therapeutic psychedelic use (e.g. MDMA therapy) and during psychological integration work. Stretching a little, perhaps we could also apply Atasoy's CSHW algorithm to *individual* brain regions and compare their spectrums (and those of CSHWs), to quantify the expected frequency-coupling between each region.[4] Possibly these two measures could be developed into a *causal quantitative metric for trauma*.

This generalized 'breakdown of communication' definition of trauma neatly fits with the story Scott tells about PTSD, where people

> *[A]re probably still "having" "emotions" "under" "the" "surface", but they have no perception of them ... PTSD patients as disconnected from their bodies. The body sends a rich flow of information to the brain, which is part of what we mean when we say we "feel alive" or "feel like I'm in my body". In PTSD, this flow gets interrupted. People feel "like nothing".*

It also fits with the therapies that seem to work: EMDR, neurofeedback, Internal Family Systems (IFS), yoga, massage — the consistent thread that connects these is they all plausibly help restart and strengthen communication within the brain (which I hold is strongly mediated by CSHWs). Scott doesn't mention music, but I'd expect it to be *surprisingly* effective at boosting emotional integration — and I'd expect the *most effective* music will have strong low-frequency rhythms.

This shades into novel types of therapeutic approaches: perhaps we could simply pump energy into lower-frequency bands (perhaps harmonic stimulation centered at ~3-6hz) to kickstart emotional integration.[5]

Sidenote on music: The simple description I gave of music was

> *[A] sensory input which seems to exist on the knife's edge between exhibiting highly ordered patterns (some of which will hit natural connectome harmonics and so allow accumulation of energy through resonance) on one hand, and on the other hand not being toopredictable (thus dodging most inhibitory top-down predictive models).*

Armed with the CSHW/RSHW distinction, we can give this a second pass. In short, I expect the above story to be true, but in a fractal way: music will be hitting both CSHWs and RSHWs. Naturally, different regions will have different sets of harmonics, which means simple tones are unlikely to produce much cross-regional resonance. Instead, the music which is the most effective at increasing the brain's energy parameter will tie together and layer a diverse set of motifs, with two goals: (1) hitting as many connectome-specific *and* region-specific resonances as possible, and (2) entraining disparate regions and pulling them into sync – essentially using injection-locking to pull RKSs (Regional Key Signatures) into sync with each other and the CKS (Connectome Key Signature).

Could we quantify what the 'perfect song' would be, for a given connectome? Not exactly, since so much of music's effects rely on getting through the brain's predictive processing gauntlet and the state of this gauntlet isn't well-captured by a static connectome, but we could possibly use this framework to design (potentially *much*) more evocative songs.

It's also worth noting that better music – and *better ways to listen to music* – shade quickly into potential therapies for trauma under this model.

V. On psychedelics:

As noted above, Neural Annealing suggests a very simple model for understanding the effects of psychedelics: as substances which "may function by disabling existing energy sinks (or perhaps overloading them by increasing baseline firing rates or increasing the branching factor of neural activity)," dramatically increasing semantically-neutral energy. Psychedelics share a 'characteristic feeling' (and

characteristic emotional aftereffects) with each other and with activities such as meditation, listening to music, EMDR, breath work, and so on, *because* all of these things increase the energy parameter of the brain. Psychedelics are particularly interesting because they do this so powerfully, effortlessly, and noisily (with the effects bleeding over into sensory modalities, not just accumulating in harmonics).

A full Neural Annealing model of psychedelics will have to wait a few more months as internal QRI discussion settles on a unified story. But a few preliminary notes:

First, we could define 'psychedelics' in a principled way, as *any substance, pattern, or process that produces semantically-neutral energy accumulation* – anything that disables, overloads, or avoids the brain's energy normalization system. The implication here is interesting, that anything that adds semantically neutral energy into the brain should produce psychedelic effects, regardless of how this is done. E.g., even things like *modern art* may be classifiable as a psychedelic, insofar as it generates semantically-neutral energy (see Gomez Emilsson 2019). But we should also note that current psychedelics are not necessarily perfect sources of 'clean semantically-neutral energy'; they're substances that happen to massively increase the energy parameter of the brain, with no guarantees about how 'balanced' this boost is. There may be better and more targeted methods to do this in the future. In the meantime, I would recommend modest caution with substances which involve a hangover after use, as negative valence or affective blunting during a critical window could 'sour' the annealing process with subtle long-term mood effects.[6]

As mentioned above, I've been thinking more and more that the core psychological changes driven by psychedelics are best understood in terms of the amount and 'statistical flavor' of the semantically-neutral energy they add to the system. Or, as an alternate framing, psychedelics may be best understood as temporary disrupters of the brain's natural energy sinks, each with a specific target or 'flavor' of disruption (or psychedelics may add to neural activity's 'branching factor', which in turn will add a specific flavor to the energy). I also find myself wondering, all else being equal, whether psychedelic visuals actually are *inversely correlated* with annealing effects, since by diverting energy into the visual system (which plausibly has very effective energy sinks), there is less energy available to drive entropic disintegration.[7]

As I noted in A Future for Neuroscience, another starting point for sorting through psychoactive drugs would be

> *[T]o parametrize the effects (and 'phenomenological texture') of all psychoactive drugs in terms of their effects on the consonance, dissonance, and noise of a brain, both in overall terms and within different frequency bands (Gomez Emilsson 2017).*
>
> *In the long term, we'll want to move upstream and predict connectome-specific effects of drugs – treating psychoactive substances as operators on neuroacoustic properties, which produce region-by-region changes in how waves propagate in the brain (and thus different people will respond differently to a drug, because these sorts of changes will generate different types of results across different connectomes). Essentially, this would involve evaluating how various drugs change the internal parameters of the CSHW model, instead of just the outputs. Moving upstream like this might be necessary to predict why e.g. some people respond well to a given SSRI, while others don't (nobody has a clue how this works right now).*

Possibly this would allow us to generate a principled typology of psychoactives, and also check for missing quadrants: psychoactives and psychedelics we haven't discovered or created yet. (See also Andrés's notion of parametrizing the 'information vs energy trajectory' of a trip.) We can also think of anti-psychotic drugs as *anti-psychedelics*: substances that *rapidly decrease the energy parameter of the brain* (Gomez Emilsson 2019). We at QRI strongly believe this makes anti-psychotics more dangerous than commonly realized: the neural search process is complex and delicate, and an externally-forced, uneven rapid cooling process may warp the internal landscape of the brain in subtle but deleterious ways. In theory, we could test this indirectly by evaluating the effects of anti-psychotics on sensory integration tasks in healthy controls – but as noted above, this may be an unethical experiment.

Another frame would be 'psychedelics as full-spectrum resonance agents' – CSHWs are meant to substantially resonate during normal human operation (falling in love, orgasm, etc) – RSHWs are not. The perceptual and epistemological changes we sometimes see during psychedelics could be due to the fine logical machinery that usually deals with high-context sensory particulars (facts and logical inferences) starting to malfunction as its natural eigenmodes are activated. Like linking and rhythmically flipping all the bits in a memory register, ignoring what that register is "supposed to" compute. If psychedelic visuals are an example of RSHW resonance, HPPD may be an example of this RSHW resonance annealing into durable patterns.

On MDMA's strangely powerful therapeutic effects, I'd suggest MDMA shares the 'basic psychedelic package' with substances like LSD and psilocybin (albeit a little weaker at common doses). Anything with this 'baseline' package significantly increases the energy parameter of the brain, which both allows escape from bad local minima and canalizes the brain's core CSHWs, which both should be highly therapeutic. My intuition is MDMA *may also* have a particular effect on stochastic firing frequencies of neurons, and that this effect essentially acts as an emergent metronome – and this metronome will drive synchronicity between diverse brain regions. Given the presence of such a region-spanning 'clean' metronomic signal, brain regions that have partially 'stopped talking to each other' will re-establish integration, and some of this integration will persist while sober (or rather, some of the reasons for the lack of integration will have been negotiated away during the MDMA-driven integration). Plausibly this 'emergent metronome' effect may also underlie the particular phenomenological effects of 5-MeO-DMT as well, particularly in terms of sense of unity, high valence, and therapeutic potential.[8]

Somewhat poetic sidenote: on taking psychedelics:

In the abstract – I think psychedelics are more powerful, more dangerous, and more healing than commonly assumed.

But we don't live in the abstract. The natural question for any given person is thus: *should I take them?*

There's no one-size-fits-all answer, and I recommend checking with local laws. But I can share a simple heuristic for who shouldn't worry too much about the downsides of psychedelics and who should be very careful: do you trust your own aesthetic?

Psychedelics massively increase the 'energy parameter' of the brain, so naturally there's a large amount of very-high-dimensional exploration going on. There are countless 'micro-choices' your brain makes as to how to anneal after this exploration: we can think of a person's 'aesthetic' as individual variance in these annealing

choices. What the self-organizing system which is the brain's subconscious finds beautiful in the moment and implicitly strives to save.

Sometimes, and in some people, we want the right things, we find the right things beautiful. Things that have a deep elegance and fit with everything about us and fit with how reality works. We just need enough energy parameter to get there. Psychedelics are a great way to get there.

Other times, we might not want the right things. Evolution is kind of a jerk, epistemologically speaking: it cares much more about genetic reproduction than it does about deep coherence and calibration with reality and such. Sometimes we're at a functional local maxima, but we're not pointed in the right direction globally, and frankly speaking our lack of a high energy parameter is our saving grace – our inability to directly muck up our emotional landscape. Insofar as this is true – and it will be more true at certain times than others, and in certain people than others, and perhaps in certain combinations of people than others – using psychedelics to crank the energy parameter is not good for a person. Our 'Psychedelic Extrapolated Volition' (PEV) is not a healthy vector.[9]

The natural follow-up is, how do you know whether your PEV is positive or not?

Hard question, but probably good to ask your friends – group epistemology seems healthy in these cases. And in general it seems strongly preferable to err on the side of caution. You can always take that LSD tomorrow, or next week, or next year.

(But, don't be too paranoid about one trip permanently breaking your brain, either. My guess is the annealing that tends to 'stick' is that which actually finds better local minima (thankfully) – if it's an unsuccessful exploration I suspect the system can usually climb back to where it was (with some caveats).)

A separate factor is your current energy parameter and how psychedelics may increase this baseline: if you're dragging on the bottom of your energetic attractor basins, maybe a little kick could be healthy. But if you're already 'high on life' – consider skipping the LSD and MDMA. Increasing a high baseline can redline the system into exquisitely unbearable intensity.

VI. Love and other types of Neural Annealing

It's important to note that most annealing doesn't happen in a vacuum: just as "set and setting" matter quite a lot for psychedelics, and for emotional updating in general, the importance of *context* in the annealing model is hard to overstate. Much as holding a magnet close to iron as it cools can magnetize the metal, the intentional content present when entropic disintegration->annealing happens provides important constraints for which new patterns form. I propose there are four general types of neural annealing:

A. Annealing to an object or event. Annealing which is 'pointed at' something is by far the most common type. Some object, or event, or new insight makes itself known in a surprising or otherwise intensely salient way, and this pushes the brain into a high-energy state, kickstarting a self-organization process for accommodating the presence and significance of this new thing. This can involve intense positive emotion — a new romantic partner, the birth of your child, your wedding day. This sort of annealing can also be caused by trauma— getting bitten by a weird animal, social rejection, losing a close one. As I suggested in The Neuroscience of Meditation, neural annealing may offer a rather pithy description of love:

*Finally, to speculate a little about one of the deep mysteries of life, perhaps we can describe **love** as the result of a strong annealing process while under the influence of some pattern. I.e., evolution has primed us such that certain intentional objects (e.g. romantic partners) can trigger high-energy states where the brain smooths out its discontinuities/dissonances, such that given the presence of that pattern our brains are in harmony. This is obviously a two-edged sword: on one hand it heals and renews our 'cold-worked' brain circuits and unifies our minds, but also makes us dependent: the felt-sense of this intentional object becomes the key which unlocks this state. (I believe we can also anneal to archetypes instead of specific people.)*

Annealing can produce durable patterns, but isn't permanent; over time, discontinuities creep back in as the system gets 'cold-worked'. To stay in love over the long-term, a couple will need to re-anneal in the felt-presence of each other on a regular basis. From my experience, some people have a natural psychological drive toward reflexive stability here: they see their partner as the source of goodness in their lives, so naturally they work hard to keep their mind aligned on valuing them. (It's circular, but it works.) Whereas others are more self-reliant, exploratory, and restless, less prone toward these self-stable loops or annealing around external intentional objects in general. Whether or not, and within which precise contexts, someone's annealing habits fall into this 'reflexive stability attractor' might explain much about e.g. attachment style, hedonic strategy, and aesthetic trajectory.

Perhaps we can go further now, and hypothesize that 'falling in love' is a specific algorithm the brain runs, which is triggered by when the 'felt sense' of another person (a pattern distributed across RSHWs, CSHWs, and SSHWs) produces substantial systemic resonance. When this happens, and in the absence of warning signs (dissonance), a person will actively seek to fill their sensorium with this signal, which amplifies the systemic resonance (potentially to extreme levels) and further synchronizes priors and other regions into harmony with the original pattern. *As you fall in love, you literally anneal to your felt-sense of that person – you take their rhythm as yours, because your body judged it to be so*. A key which fit your connectome's lock. This will naturally do two things: (1) fuzz boundaries between lovers, as patterns progressively synchronize, and (2) add a harmonic echo, or 'warm consonant glow' to all thoughts about the person. This latter phenomenon will feel nice, but also keep itself stable: the presence of this bundle of synchronized frequencies will stabilize (via injection-locking) many forms of drift – effectively preventing certain thoughts/perceptions. This may fade over time if not refreshed, but perhaps to completely 'fall out of love' the brain has to build a competing key signature elsewhere, e.g. in a golden mean ratio to this harmonic echo, and these rivalrous key signatures (implicitly Bayesian priors about what is real and what is good) battle it out. (Thanks to Andrés for discussion on competing key signatures.) This 'de-annealing' process – literally erasing someone's patterns and rhythms from your body – can follow several trajectories, few of them pleasant, as the system renegotiates new (or old) equilibria.

<u>B. Annealing to an ontology</u>. A much more general type of annealing is when the entropic disintegration->annealing process is pointed toward an *ontology*, and the brain reorganizes its internal structure ('ontological contours') to accommodate this new typology. This can happen implicitly and weakly, over the course of entropic disintegration->annealing to multiple separate ideas, or explicitly and strongly, for instance reading some book in college which completely reshapes one's view of reality.

Any craftsman, any intellectual, any philosopher worth their salt is strongly annealed toward at least one nuanced ontology, and in fact much of the influence of the Great Philosophers can be found in how they've laid out their thoughts in a way that others can use as a *coherent annealing target*. What makes something a good annealing target? I'd offer it's the presence of clear archetypes arranged in both a novel but ultimately cognitively efficient way. These archetypes can be thought of as a combination of nature (innate Jungian-type limbic resonances) and nurture (prior annealed patterns & cultural reifications).

An important point here is that peoples' conception of *where goodness comes from* is dependent upon their ontology; change the ontology, change the perceived nature of goodness itself! See e.g. John Lily's discussion of the supra-self-metaprogrammer (SSMP). This frame-shift can also manifest at the extreme end of falling in love, where all the world's goodness seems to come from your special person (a dangerous thing).

C. Social annealing. A special hybrid of annealing to an ontology and to other people is **social annealing**, wherein a *group of people* undergoes the *'entropic disintegration -> neural search -> annealing'* process together, within some shared context- a religious service, a sporting event, a retreat. This seems like the natural mechanism by which tribes are formed (loosely speaking, group synchronization of connectome-specific harmonic wave dynamics) and underlies many of our most sacred experiences. The power of social annealing is such that a religious experience that lacks it no longer feels like a religious experience- merely the mouthing of dogma. On the other hand, any group experience that *does* increase the group's energy parameter and trigger annealing starts to take on a pseudo-religious frame- e.g. ecstatic dance, festivals, protest marches, even concerts.

D. Semantically-neutral annealing. Almost all neural annealing is *semantic annealing*, or *annealing toward some intentional object*. This process is pointed *at* something, often the thing that caused the entropic disintegration process in the first place, be it a person, an event, an idea, an ontology. But there's nothing in the laws of neuroscience that implies annealing *has to* have an intentional object as a focus. As per Section II, I believe this is a particularly healthy form of annealing.

Toward a new psychology & sociology?

Speculatively, we may be able to re-derive much of psychology and sociology from just the energy-parameter view of the brain: e.g.,

Gopnik 2017 suggests that different developmental windows may involve different implicit 'heat parameters' for simulated annealing, with young people having higher parameters. Speculatively, this may correspond to different 'lived intensity of experience' at different ages- young brains (and lifelong learners) might not only be *more plastic* than average, but actually having experience that is objectively *more visceral*. One way to frame this is that *being young is like microdosing on LSD all the time.* This could have interesting implications for ethics.

Most likely, there's been significant recent sexual selection for a higher energy parameter, for several reasons:

- Selecting for neoteny plausibly also implicitly selects for an energy parameter that starts higher and/or decays less with age;
- A high energy parameter would be a good proxy for cognitive-emotional-behavioral dynamicism, perhaps the most strongly sexually-selected-for trait;

- A high energy parameter would be an honest signal of not being in a bad 'iterated aesthetics' attractor (otherwise they would have self-destructed previously).

Psychology has various personality metrics, with the most widely used being the Big 5, also known as OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism). One of the most interesting subfindings here is that we can still get reasonable predictive utility if we collapse these into a one-variable model: the 'Big One' personality factor. Scoring high in this factor is "associated with social desirability, emotionality, motivation, well-being, satisfaction with life, and self-esteem." Scoring low is associated with depression, frailty, lack of emotionality, and so on. I wouldn't be surprised if the 'Big One' simply tracks how frequently and deeply someone anneals.

Continuing the thread on Social Annealing, I think we can push into sociology with the Neural Annealing model too; **to understand a society, we need to understand how and when annealing happens in that society**. To gauge the wisdom of a society, look at how its decision-makers anneal; to gauge the cultural direction of a society, look at how its young people anneal. To understand the strongest social bonds of a society, look at the contexts in which group annealing happens.

This also suggests why drugs like alcohol and certain psychedelics are ritualistically celebrated in so many cultures: they allow social-annealing-on-demand, a key technology in building and maintaining social cohesion and coordination.

Likewise, we could envision a field of 'social archeology' evaluating annealing patterns in the past: how often did peasants and nobles in the Middle Ages anneal? In what contexts did the annealing happen, and which institutions controlled them? Perhaps most political conflicts could be reinterpreted as conflicts over annealing.[10] And so on. My colleague Andrés has suggested that a good rule of thumb for identifying annealing (and making good movies) is that intuitively, annealing defines where you should actually point the camera if you were making a movie of a historical period, since where annealing is happening is where changes that 'matter' are taking place: cognitive updates, decisions about how to feel, and so on.

On the effect of profession on emotional vibrancy: It would be somewhat surprising if certain repeated computational tasks didn't tend to push regions' key signatures into being highly coupled (=intense emotions), whereas other classes of tasks push regions' key signatures into fairly orthogonal configurations (='white noise' as emotional state). A lifetime of dance or poetry might literally make you feel emotions more strongly; a lifetime of doing accounting might literally produce a segmented brain and affective blunting. From Darwin's autobiography:

> *I have said that in one respect my mind has changed during the last twenty or thirty years. Up to the age of thirty, or beyond it, poetry of many kinds, such as the works of Milton, Gray, Byron, Wordsworth, Coleridge, and Shelley, gave me great pleasure, and even as a schoolboy I took intense delight in Shakespeare, especially in the historical plays. I have also said that formerly pictures gave me considerable, and music very great delight. But now for many years I cannot endure to read a line of poetry: I have tried lately to read Shakespeare, and found it so intolerably dull that it nauseated me. I have also almost lost my taste for pictures or music. … I retain some taste for fine scenery, but it does not cause me the exquisite delight which it formerly did. …*

> *This curious and lamentable loss of the higher aesthetic tastes is all the odder, as books on history, biographies, and travels (independently of any scientific facts which they may contain), and essays on all sorts of subjects interest me as much as ever they did. My mind seems to have become a kind of machine for grinding general laws out of large collections of facts, but why this should have caused the atrophy of that part of the brain alone, on which the higher tastes depend, I cannot conceive.*

Reading this account, I find it plausible that Darwin repeatedly pushed (and annealed) his mind toward RSHW-driven 'clockwork piecemeal integration' interactions rather than CSHW/SSHW-driven global symmetry gradients, although Darwin's age, sickness, and depression may have also contributed. A warning sign for us theorists and systematizers.

Conclusion:

Neural Annealing is a neuroscience paradigm which aims to find the optimal tradeoff between elegance and detail. It does this by identifying a level of abstraction which supports parallel description under three core principles of self-organization: physical self-organization (around connectome resonances), computational self-organization (around minimization of surprise), and energetic self-organization (around conditional entropic disintegration).

There is yet much work to be done: in particular, there are huge bodies of literature around receptor affinities, network topologies, regional anatomies and cell types, and so on. The promise of Neural Annealing is it's not only a predictive and generative theory in its own right, but it provides a level of description by which to connect these disparate maps, and an extensible context to build on as we add more and more detail to the model.

Finally, we can ask: why does good neuroscience *matter*? I would offer the following.

The future could be much better than the present. *Much* better.

Material conditions are only very loosely coupled with well-being. If life is to be radically better in the future, it will be due to better neuroscience pointing out how we can be kinder to ourselves and others, and future neurotechnology changing the hedonic calculus of the human condition.

A unified theory of emotional updating, depression, trauma, meditation, and psychedelics may give us the tools to build a future that's *substantially* better than the present. This has been my hope while writing this.

–Michael Edward Johnson, Executive Director, Qualia Research Institute

Endnotes:

[1] The 'semantically neutral energy' model also suggests why transcranial magnetic stimulation (TMS) seems to help treat depression – essentially, TMS injects a large amount of energy into the brain, and this energy (1) triggers some entropic disintegration, allowing escape from bad local minima, and (2) may *slightly* collect in the brain's natural harmonics, which may help pull the brain out of dissonant equilibria. Note that this could be done *much more effectively*: instead of the present strategy of using a quick flash of unpatterned, pulsed TMS (e.g., 5 seconds @ 100hz) which overpowers the brain but quickly dissipates and likely doesn't lead to a

significant build-up in harmonics, we could instead try an entrainment approach via lower-power, rhythmic, continuous TMS, applied for longer durations (keeping the brain above its 'recrystallization temperature' for longer, allowing a fuller self-organization process), perhaps paired with music.

[2] Thanks to Andrés for the idea about somatic information, and the suggestion of sensorium as the label.

[3] I suspect that *muscle tension* could be a core mechanism for regulating SSHWs and perhaps CSHWs. Tensing muscles will strongly influence body resonance, and one's body resonance configuration will likely have ripple effects on what sorts of frequencies persist in the brain. This suggests that traditions such as yoga are basically right when they posit a link between problems in muscles and problems in the mind: we may hold tension in one system in order to compensate for a problem in the other. Speculatively, this compensatory regulation may also be found *across* humans, especially in pair bonds: that tension in your back might in some literal way be an attempt to help your partner with their emotional regulation. This would suggest muscle tension should change significantly after a break-up. (Thanks to Emily Crotteau, Lena Selesneva, and Ivanna Evtukhova for pieces of the puzzle here.)

[4] My colleague Andrés suggests that "[A] more direct method, though perhaps more difficult, would be to look directly for the spectral signatures of injection locking — we'd predict you will see a seriously diminished degree of injection locking signatures on people who are heavily traumatized, and see it come back after MDMA therapy."

[5] Perhaps we could model Persistent Non-Symbolic Experience (PNSE) as persistent partial injection locking of key regions by low frequency CSHWs: essentially this would involve entraining (and effectively partially disabling) the machinery that usually handles interpretation of certain particulars / cognitive interpretations. Perhaps highly neurotic or traumatized individuals with strong top-down control exhibit the opposite: essentially trying to entrain CSHWs to a specific region (with predictably poor results).

[6] My colleague Andres also recommends against "psychedelic substances that have as part of their activity profile a high level of body-load, such as nausea and cramps as these patterns might themselves become annealing targets (cf. compounds notorious for this, according to PsychonautWiki such as 2C-E, 2C-T-2, and 2C-P, are probably best avoided as therapeutic aids)."

[7] On psychedelic tolerance: if the semantically-neutral energy model of psychedelics proves out, we should also be open to subtle corollaries: e.g., to what extent is the temporary tolerance effect of psychedelics *biochemical* (depletion of some neurotransmitters, per the current story) and to what extent is it *information-theoretic* —associated with the release and depletion of systemic sources of Free Energy? I.e., there is potential energy of a sort liberated when the system finds a better local minima, and if the system has undergone strong annealing recently, there are fewer such 'energetic free lunches' around to help power the psychedelic effects. (Hypothesis held weakly, as my colleague Andrés points out there are psychedelics which do not trigger tolerance, such as N,N-DMT and 5-MeO-DMT.)

[8] HT to Steve Lehar for pointing at this 'nystagmus' phenomenon as being somehow linked to MDMA's mood-lifting effect, and to Andrés for calling my attention to Lehar's work and suggesting 5-MeO-DMT may also share this mechanism.

[9] This is a reference to Eliezer Yudkowsky's "Coherent Extrapolated Volition" (CEV) concept, which is an attempt to sketch a heuristic for how to use a radically-powerful

optimization process (such as an AGI) safely. Essentially, CEV suggests we could aggregate all human preferences (volitions), find some way to merge them (make them 'cohere'), then repeat (extrapolate), until we get to a self-stable loop. A 'psychedelic extrapolated volition' is a variation of this: if it becomes easier to change yourself on psychedelics, and then that person you turn into can change themselves into someone else, and so on, where do you end up? What generates a 'positive vector' here?

[10] This naturally and unfortunately makes the access to and contexts of social annealing an axis of cultural conflict: those who control these events control the emotional tone and contours of coordination of a society. Taking away healthy annealing contexts from your opponents and giving more social annealing opportunities to your people is a key (but also very dirty) way to 'win' the culture war. (Perhaps the opioid crisis, and the crack-cocaine crisis before it, could in some sense be exacerbated by a lack of healthier annealing opportunities.)

Citation

For attribution in academic contexts, please cite this work as:

Michael Edward Johnson, "Neural Annealing: Toward a Neural Theory of Everything", https://opentheory.net/2019/11/neural-annealing-toward-a-neural-theory-of-everything/ , San Francisco (2019).

Acknowledgements

I'd like to thank Andrés Gómez Emilsson for many great conversations on annealing (and first calling my attention to the term), energy sinks, and countless other topics, and offering careful feedback on a draft of this work; Robin Carhart-Harris and Karl Friston for a beautiful description of simulated annealing; Romeo Stevens for wide discussion about annealing & ontologies; Adam Safron for introducing me to the depth of explanation afforded by predictive coding, pointing me toward injection locking, and many great conversations in general; Quintin Frerichs for his hard work toward making therapeutic applications of this theory real, and the rest of the QRI team for support and inspiration; Milan Griffes for careful feedback on a draft of this work; Alex Alekseyenko and James Dama for discussions about simulated annealing; Anthony Markwell for sharing the Buddhist Dhamma with me in such a thoughtful and generous way; Justin Mares for his constant curiosity and encouragement; my parents, for their endless love and patience; Lena Zaitseva and Lena Selesneva for their warmth and support; and especially Ivanna Evtukhova, who has made my life radically better and whose love, energy, and obsession with Buddhist enlightenment was why this work happened.

To gratitude.

Timeline: most of this document written ~Feb-April 2019, as a continuation of The Neuroscience of Meditation and this talk, and shared internally and with select reviewers; section dealing with trauma written July 2019. Document reordered for flow and polished in Oct-Nov and posted Thanksgiving 2019.

Historical forecasting: Are there ways I can get lots of data, but only up to a certain date?

Suppose I wanted to get good intuitions about how the world works on historical timescales.

I could study history, but just reading history is rife with historical hindsight bias, both on my own part, and even worse, on the part of the authors I'm reading.

So if I wanted to master history, a better way would be to do it forecasting-style. I read what was happening in the some part of the world, up to a particular point in time, and then make bets about what will happen next. This way, I have feedback as I'm learning, and I'm training an actual historical predictor.

However, this requires a **strong** limit be enforced on the materials I'm reading: no information about "what's going to happen" can leak backwards. And unfortunately, this is kind of standard in history books. Usually, the author talk about how events are leading towards other events that they know will occur.

Is there some databases (or something), where I might be able to read a wide number of primary sources and economic / socioeconomic indicators (like the amount of pottery fragments, average skeleton size, how far specialized goods traveled, how much money was in circulation, the literacy rate, etc.), but which will only show me data **up to** a certain date, with a strong constraint of not accidentally seeing spoilers?

Epistemic Spot Check: Unconditional Parenting

Epistemic spot checks started as a process in which I investigate a few of a book's claims to see if it is trustworthy before continuing to read it. This had a number of problems, such as emphasizing a trust/don't trust binary over model building, and emphasizing provability over importance. I'm in the middle of revamping ESCs to become something better. This post is both a ~ESC of a particular book and a reflection on the process of doing ESCs and what I have and should improve(d).

As is my new custom, I took my notes in Roam, a workflowy/wiki hybrid. Roam is so magic that my raw notes are better formatted there than I could ever hope to make them in a linear document like this, so I'm just going to share my conclusions here, and if you're interested in the process, follow the links to Roam. Notes are formatted as follows:

- The target source gets its own page
- On this page I list some details about the book and claims it makes. If the claim is citing another source, I may include a link to the source.
- If I investigate a claim or have an opinion so strong it doesn't seem worth verifying ("Parenting is hard"), I'll mark it with a credence slider. The meaning of each credence will eventually be explained here, although I'm still working out the system.
    - Then I'll hand-type a number for the credence in a bullet point, because sliders are changeable even by people who otherwise have only read privileges.
- You can see my notes on the source for a claim by clicking on the source in the claim
- You may see a number to the side of a claim. That means it's been cited by another page. It is likely a synthesis page, where I have drawn a conclusion from a variety of sources.

This post's topic is Unconditional Parenting (Alfie Kohn) (affiliate link), which has the thesis that even positive reinforcement is treating your kid like a dog and hinders their emotional and moral development.

Unconditional Parenting failed its spot check pretty hard. Of three citations I actually researched (as opposed to agreed with without investigation, such as "Parenting is hard"), two barely mentioned the thing they were cited for as an evidence-free aside, and one reported exactly what UP claimed but was too small and subdivided to prove anything.

Nonetheless, I thought UP might have good ideas kept reading it. One of the things Epistemic Spot Checks were designed to detect was "science washing"- the process of taking the thing you already believe and hunting for things to cite that could plausibly support it to make your process look more rigorous. And they do pretty well at that. The problem is that science washing doesn't prove an idea is wrong, merely that it hasn't presented a particular form of proof. It could still be true or useful- in fact when I dug into a series of self-help books, rigor didn't seem to have any correlation with how useful they were. And with something like child-rearing, where I dismiss almost all studies as "too small, too limited", saying everything needs rigorous peer-reviewed

backing is the same as refusing to learn. So I continued with *Unconditional Parenting* to absorb its models, with the understanding that I would be evaluating its models for myself.

Unconditional Parenting is a principle based book, and its principles are:

- It is not enough for you to love your children; they must feel loved unconditionally.
- Any punishment or conditionality of rewards endangers that feeling of being loved unconditionally.
- Children should be respected as autonomous beings.
- Obedience is often a sign of insecurity.
- The way kids learn to make good decisions is by making decisions, not by following directions.

These seem like plausible principles to me, especially the first and last ones. They are, however, costly principles to implement. And I'm not even talking about things where you absolutely have to override their autonomy like vaccines. I'm talking about when your two children's autonomies lead them in opposite directions at the beach, or you will lose your job if you don't keep them on a certain schedule in the morning and their intrinsic desire is to watch the water drip from the faucet for 10 minutes.

What I would really have liked is for this book to spend less time on its principles and bullshit scientific citations, and more time going through concrete real world examples where multiple principles are competing. Kohn explicitly declines to do this, saying specifics are too hard and scripts embody the rigid, unresponsive parenting he's railing against, but I think that's a cop out. Teaching principles in isolation is easy and pointless: the meaningful part is what you do when they're difficult and in conflict with other things you value.

So overall, *Unconditional Parenting*:

- Should be evaluated as one dude's opinion, not the outcome of a scientific process
- Is a useful set of opinions that I find plausible and intend to apply with modifications to my potential kids.
- Failed to do the hard work of demonstrating implementation of its principles.
- Is a very light read once you ignore all the science-washing.

As always, tremendous thanks to my Patreon patrons for their support.

A test for symbol grounding methods: true zero-sum games

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

Imagine there are two AIs playing a debate game. The game is zero-sum; at the end of the debate, the human judge assigns the winner, and that AI gets a $+1$ reward, while the other one gets a -1.

Except the game, as described, is not truly zero-sum. That is because the AI "get" a reward. How is that reward assigned? Presumably there is some automated system that, when the human presses a button, routes $+1$ to one AI and -1 to another. These rewards are stored as bits, somewhere "in" or around the two AIs.

Thus there are non zero-sum options: you could break into the whole network, gain control of the automated system, and route $+1$ to each AI - or, why not, $+10^{100}$ or even $+f_{\psi(\Omega^{\Omega^{\Omega}})}(4)$ or whatnot[1].

Thus, though we can informally say that "the AIs are in a zero-sum game as to which one wins the debate", that sentence is not properly grounded in the world; it is only true as long as certain physical features of the world are maintained, features which are not mentioned in that sentence.

Symbol grounding implies possibility of zero-sum

Conversely, imagine that an AI has a utility/reward U/R which is properly grounded in the world. Then it seems that we should be able to construct an AI with utility/reward $-U/-R$ which is also properly grounded in the world. So it seems that any good symbol grounding system should allow us to define truly zero sum games between AIs.

There are, of course, a few caveats. Aumann's agreement theorem requires unboundedly rational agents with common priors. Similarly, though properly grounded U and $-U$ are zero-sum, the agents might not be fully zero-sum with each other, due to bounded rationality or different priors.

Indeed, it is possible to setup a situation where even unboundedly rational agents with common prior will knowingly behave in not-exactly zero-sum ways with each other; for example, you can isolate the two agents from each other, and feed them deliberately biased information.

But those caveats aside, it seems that proper symbol grounding implies that you can construct agents that are truly zero-sum towards each other.

Zero-sum implies symbols grounded?

Is this an equivalence? If two agents really do have zero sum utility or reward functions towards each other, does it mean that those functions are well grounded[2]?

It seems that it should be the case. Zero-sum between U and $V = -U$ means that, for all possible worlds w, $U(w) = -V(w)$. There are no actions that we - or any agent - could do that breaks that fundamental equality. So it seems that U must be defined by features of the world; grounded symbols.

Now, these grounded symbols might not be exactly what we thought they were; its possible we thought U was defined on human happiness, but it is actually only means current in a wire. Still, V must then be defined in terms of absence of current in the wire. And, whatever we do with the wire - cut it, replace it, modify it in cunning ways - U and V must reach opposite on that.

Thus it seems that either there is *some* grounded concept that U and V are opposite on, or U and V contain exhaustive lists of all special cases. If we further assume that U and V are not absurdly complicated (in a "more complicated than the universe" way), we can rule out the exhaustive list.

So, while I can't say with full confidence that a true zero-sum game must mean that the utilities are grounded, I would take such a thing as a strong indication that they are.

---

1. If you thought that $3 \uparrow\uparrow\uparrow 3$ was large, nothing will prepare you for $f_{\psi(\Omega^{\Omega^{\Omega}})}(4)$ - the fast-growing hierarchy indexed by the large Veblen Ordinal. There is no real way to describe how inconceivably huge this number is. ↩
2. Assuming the functions are defined in the world to some extent, not over platonic mathematical facts. ↩

Platonic rewards, reward features, and rewards as information

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

Contrast these two expressions (hideously mashing C++ and pseudo-code):

1. $\text{argmax}_x\, r(x)$,
2. $\text{argmax}_x\, *(\&r)(x)$.

The first expression just selects the action x that maximises r(x) for some function r(), intended to be seen as a reward function.

The second expression borrows from the syntax of C++; (&r) means the memory address of r, while $*(\&r)$ means the object at the memory address of r. How is that different from r itself? Well, it's meant to emphasise the ease of the agent wireheading in that scenario: all it has to do is overwrite whatever is written at memory location (&r). Then $*(\&r)$ can become - whatever the agent wants it to be.

Let's dig a bit deeper into the contrast between reward functions that can be easily wireheaded and those that can't.

The setup

The agent A interacts with the environment in a series of timesteps, ending at time $t = N$.

There is a 'reward box' R which takes observations/inputs o_t^R and outputs some numerical reward amount, given by the voltage, say. The reward function is a function of o_t^R; at timestep t, that function is $R_t()$. The reward box will thus give out a reward of $R_t(o_t^R)$. Initially, the reward function that R() implements is $r() = R_0()$.

The agent also gets a separate set of observations o_t^A; these observations may include full information about o_t^R, but need not.

Extending the training distribution

Assume that the agent has had a training phase, for negative values of t. And, during that training phase, $R_t()$ was always equal to r().

If A is trained as a reinforcement agent, then there are two separate value functions that it A can learn to maximise:

1. $E\sum_{t=0}^{N} r(o_t^R)$, or
2. $E\sum_{t=0}^{N} R_t(o_t^R)$.

Since $R_t() = r()$ for $t < 0$, which is all the t that the agent has ever seen, both fit the data. The agent has not encountered a situation where it can change the physical behaviour of R() to anything other than r() - how will it deal with that?

Wireheading is in the eye of the beholder

Now, it's tempting to call $r(o_t^R)$ the true reward, and $R_t(o_t^R)$ the wireheaded reward, as the agent redefines the reward function in the reward box.

But that's a judgement call on our part. Suppose that we wanted the agent to maintain a high voltage coming out of R, to power a device. Then $R_t(o_t^R)$ is the 'true' reward, while r() is just information about what the current design of R is.

This is a key insight, and a reason that avoiding wireheading is so hard. 'Wireheading' is not some ontologically fundamental category. It is a judgement on our part: some ways of increasing a reward are legitimate, some are not.

If the agent is to ever agree with our judgement, then we need ways of getting that judgement into the agent, somehow.

Solutions

The agent with a platonic reward

One way of getting the reward into the agent is to formally define it within the agent itself. If the agent knows r(), and is explicitly trained to maximise it, then it doesn't matter what R changes to - the agent will still be wanting to maximise r(). Indeed, in this situation, the reward box R is redundant, except maybe as a training crutch.

What about self modification, or the agent just rewriting the observations o_t^R? Well, if the agent is motivated to maximise the sum of r(), then, by the same argument as the standard Omohundro "basic AI drives", it will want to preserve maximising r() as the goal of its future copies. If we're still worried about this, we could insist that the agent knows about its own code, including r(), and acts to preserve r() in the future[1].

The agent modelling the reward

Another approach is to have the agent treat the data from its training phase as information, and try to model the correct reward.

This doesn't require the agent to have access to r(), from the beginning; the physical R provides data about, and after the training phase, the agent might disassemble R entirely (thus making R_t trivial) in order to get at what r() is.

This approach relies on the agent having good priors over the reward function[2], and on these priors being well-grounded. For example, if the agent opens the R box and sees a wire splitting at a certain point, it has to be able to interpret this a data on r() in a sensible way. In other words, we interpret R as implementing a function between o_t^R and the reward signal; it's important that the agent, upon disassembling or analysing R, interprets it in the same way.

Thus we need the agent to interpret certain features of the environment (eg the wire splitting) in the way that we want it to. We thus want the agent to have an model of the environment, where the key features are abstracted in the correct way.

Provide examples of difference

The third example is to provide examples of divergence between r() and $R_t()$. We could, for example, unplug the R for a time, then manually give the agent the rewards during that time. This distinguishes the platonic $r(o_t^R)$ we want it to calculate, from the physical $R_t(o_t^R)$ that the reward function implements.

Modifying the $R_t()$ during the training phase, while preserving r() as the "true" reward, also allows the agent to learn the difference. Since $R_t()$ is physical while r() is idealised, it might be a good idea to have the agent explicitly expect noise in the $R_t(o_t^R)$ signal; including the expectation of noise in the agent's algorithm will prevent the agent overfitting to the actual $R_t()$, increasing the probability that it will figure out the idealised r().

In the subsection above, we treated features of the reward box as information that the agent had to interpret in the desirable way. Here we are also wanting the agent to learn features correctly, but we are teaching it by example rather than by model - exploring the contours of the feature, illustrating when the feature behaved in a proper informative way, and when it failed to do so.

Extending the problem

All the methods above suffer from the same issue: though they might prevent wireheading *within* r() or R, they don't prevent wireheading *at the input boundary* for R, namely the value of o_t^R.

As it stands, even the agent with a platonic reward will be motivated to maximise r() by taking control of the values of o_t^R.

But the methods that we have mentioned above allow us to (try and) extend the boundary of wireheading out into the world, and hopefully reduce the problem.

To model this, let W_t be the world at time t (which is a function of the agent's past actions as well as the world's past), and let F be the set of functions f such that $f(W_t) = o^R_t$ for $t < 0$.

Similarly to how the agent can't distinguish between $R_t()$ and r(), if they were equal for all $t < 0$, the agent can't distinguish whether it should really maximise f_1 or f_2, if they are both in F.

In a sense, the agent's data for $t < 0$ allows it to do function approximation for its objective when $t \geq 0$. And we want it to learn a good function, not a bad, wireheady one (again, the definition of wireheading depends on our objective).

The easiest and most general f to learn is likely to be the one that outputs the actual physical o^R_t; and this is also the most likely candidate for leading to wireheading. If we want the agent to learn a different f, how can we proceed?

Platonic

If we can model the whole world (or enough of it), we can present the agent with a labelled model of the whole environment, and maybe specify exactly what gives a true reward and what doesn't.

This is, of course, impossible; but there are variants on this idea which might work. In these examples, the agent is running on a virtual environment, and knows the properties of this virtual environment. It is then motivated to achieve goals within that virtual environment, but the second that the virtual environment doesn't behave as expected - the second our abstraction about that environment fails - the agent's motivation changes.

Modelling true reward

It might not be feasible to provide the agent with a full internal model of the environment; but we might be able to do part of the job. If we can give the agent a grounded definition of key features of the environment - what counts as a human, what counts as a coffee cup, what counts as a bin - then we can require the agent to model its reward function as being expressed in a particular way by those features.

Again, this is controlling the abstraction level at which the agent interprets the reward signal. So, if an agent sees a human get drenched in hot coffee and fall into a bin, it

will interpret that as just described, rather than seeing it as the movement of atoms out in the world - or as the movement of electrons within its own circuits.

Examples of feature properties

Just as in the previous section, the agent could be taught about key features of the environment by example - by showing examples of good behaviour, of bad behaviour, of when abstractions hold (a human hand covered in hot coffee is still a human hand) and of when they fail (a human hand long detached from its owner is not a human, even in part, and need not be treated in the same way).

There are a lot of ideas as to who these principles could be illustrated, and, the more complex the environment and the more powerful the agent, the less likely it is that we have covered all the key examples sufficiently.

Feature information

Humans will not be able to describe all the properties of the key features we want to have as part of the reward function. But this is an area where the agent can ask the humans for more details about them, and get more information. Does it matter if the reward comes as a signal immediately, a delayed signal, or as a letter three weeks from now? When bringing the coffee to the human, what counts as spilling it, and what doesn't? Is this behaviour ok? What about this one?

Unlike issues of values, where its very easy to get humans confused and uncertain by expanding to new situations, human understanding of the implicit properties of features is more robust - the agent can ask many more questions, and consider many more hypotheticals, without the web of connotations of the feature collapsing.

This one way we could combat Goodhart problems, by including all the uncertainty and knowledge. In this case, the agent's definition of the key property of the key features, is... what a human could have told it about them if it had asked.

---

1. This is an area where there is an overlap between issues of wireheading, symbol grounding, and self-modelling. Roughly speaking, we want a well-specified, well-grounded reward function, and an agent that can model itself in the world (including knowing the purpose of its various components), and that can distinguish which features of the world it can legitimately change to increase the reward, and which features it should not change to increase reward.

   So when an agent misbehaves with a "make humans happy" reward, this might be because terms like "humans" and "happy" are incorrectly defined, or not well-grounded, or because the agent wireheads the reward definition. In practice, there is a lot of overlap between all these failure modes, and they cannot necessarily be cleanly distinguished. ↩

2. The platonic case can be seen as modelling, with a trivial prior. ↩

The Bus Ticket Theory of Genius

This is a linkpost for http://www.paulgraham.com/genius.html

Thoughts on Robin Hanson's AI Impacts interview

There was already a LessWrong Post here. I started writing this as a comment there, but it got really long, so here we are! For convenience, here is the link to interview transcript and audio, in which he argues that AGI risks are modest, and that EAs spend too much time thinking about AGI. I found it very interesting and highly recommend reading / listening to it.

That said, I disagree with almost all of it. I'm going to list areas where my intuitions seem to differ from Robin's, and where I'm coming from. Needless to say, I only speak for myself, I'm not super confident about any of this, and I offer this in the spirit of "brainstorming conversation" rather than "rebuttal".

How likely is it that the transition to superhuman AGI will be overwhelmingly important for the far future?

Robin implies that the likelihood is low: "How about a book that has a whole bunch of other scenarios, one of which is AI risk which takes one chapter out of 20, and 19 other chapters on other scenarios?" I find this confusing. What are the other 19 chapter titles? See, in my mind, the main categories are that (1) technological development halts forever, or (2) AGI is overwhelmingly important for the far future, being central to everything that people and societies do (both good and bad) thereafter. I don't immediately see any plausible scenario outside of those two categories ... and of those two categories, I put most of the probability weight in (2).

I assume Robin would want one of the 20 chapters to be about whole-brain emulation (since he wrote a whole book about that), but even if whole-brain emulation happens (which I think very unlikely), I would still expect fully-artificial intelligence to be overwhelmingly important in this scenario, as soon as the emulations invent it—i.e. this would be in category 2. So anyway, if I wrote a book like that, I would spend most of the chapters talking about AGI risks, AGI opportunities, and what might happen in a post-AGI world. The rest of the chapters would include things like nuclear winter or plagues that destroy our technological civilization forever. Again, I'm curious what else Robin has in mind.

How hard is it to make progress on AGI safety now? How easy will it be in the future?

I could list off dozens of specific open research problems in AGI safety where (1) we can make real progress right now; (2) we *are* making real progress right now; (3) it doesn't seem like the problems will resolve themselves, or even become substantially easier, after lots more research progress towards building AGI.

Here's a few off the top of my head: (1) If we wind up building AGIs using methods similar to today's deep RL, how would we ensure that they are safe and beneficial? (This is the "prosaic AGI" research program.) (2) If we wind up building AGIs using algorithms similar to the human brain's, how would we ensure that they are safe and beneficial? (3) If we want task-limited AGIs, or norm-following AGIs, or impact-limited AGIs, or interpretable AGIs, what *exactly* does this mean, in terms of a specification that we can try to design to? (4) Should we be trying to build AGI agents with explicit goals, or "helper AIs", or oracles, or "microscope AIs", or "tool AIs", or what? (5) If our AGIs have explicit goals, what should the goal be? (6) Max Tegmark's book lists 12 "AI aftermath scenarios"; what post-AGI world do we want, and what AGI research, strategy, and policies will help us get there? ...

Robin suggests that there will be far more work to do on AGI safety in the future, when we know what we're building, we're actually building it, and we have to build it right. I agree with that 100%. But I would phrase it as "even more" work to do in the future, as opposed to implying that there is not much to do right now.

How soon are high-leverage decision points?

Robin suggests that we should have a few AGI safety people on Earth, and their role should be keeping an eye on developments to learn when it's time to start real work, and that time has not yet arrived. On the contrary, I see key, high-leverage decision points swooshing by us as we speak.

The type of AI research we do today will determine the type of AGI we wind up building tomorrow; and some AGI architectures are bound to create worse safety & coordination problems than others. The sooner we establish that a long-term research program is leading towards a problematic type of AGI, the easier it is for the world to coordinate on not proceeding in that research program. On one extreme, if this problematic research program is still decades away from fruition, then not pursuing it (in favor of a different path to AGI) seems pretty feasible, once we have a good solid argument for why it's problematic. On the opposite extreme, if this research program has gotten all the way to working AGI code posted on GitHub, well good luck getting the whole world to agree not to run it!

How much warning will we have before AGI? How much do we need?

Lots of AGI safety questions seem hard (particularly, "How do we make an AGI that robustly does what we want it to do, even as it becomes arbitrarily capable and knowledgeable?", and also see the list a few paragraphs above). It's unclear what the answers will look like, indeed it's not yet proven that solutions even exist. (After all, we only have one example of an AGI, i.e. humans, and they display all sorts of bizarre

and destructive behaviors.) When we have a misbehaving AGI right in front of us, with a reproducible problem, that doesn't mean that we will know how to fix it.

Thus, I see it as entirely possible that AIs develop gradually into more and more powerful AGIs over the course of a decade or two, and with each passing year, we see worse and worse out-of-control-AGI accidents. Each time, people have lots of ideas about what the solution is, and none of them work, or the ones that work also make the AGI less effective, and so people keep experimenting with the more powerful designs. And the accidents keep getting worse. And then some countries try to regulate AGI research, while others tell themselves that if only the AGI were *even more* capable, then the safety problems would resolve themselves because the AGI would understand humans better, and hey it can even help chase down and destroy those less-competent out-of-control AGIs from last year that are still self-reproducing around the internet. And the accidents get even worse still ... and on and on...

(ETA: For more on this topic, see my later post On unfixably unsafe AGI architectures.)

This is the kind of thing I have in mind when I say that even a very gradual development of AGI poses catastrophic risks. (I'm not saying anything original here; this is really the standard argument that if AGI takes N years, and AGI safety research takes N+5 years, then we're in a bad situation ... I'm just trying to make that process more vivid.) Note that I gave an example focused on catastrophic accidents, but of course risk is disjunctive. In particular, in slow-takeoff scenarios, I often think about coordination problems / competitive pressures leading us to a post-AGI world that nobody wanted.

That said, I do also think that fast takeoff is a real possibility, i.e. that we may well get very powerful and dangerous AGI with little or no warning, as we improve learning-and-reasoning algorithms. Humans have built a lot of tools to amplify our intellectual power, and maybe "AGI code version 4" can really effectively take advantage of them, while "AGI code version 3" can't really get much out of them. By "tools" I am thinking of things like coding (recursive self-improvement, writing new modules, interfacing with preexisting software and code), taking in human knowledge (reading and deeply understanding books, videos, wikipedia, etc., a.k.a. "content overhang") , computing hardware (self-reproduction / seizing more computing power, a.k.a. "hardware overhang"), the ability of humans to coordinate and cooperate (social manipulation, earning money, etc.) and so on. It's hard to say how gradual the transition will be between not getting much out of these "tools" versus really being able to use them to their full potential, and don't see why a fast transition (weeks or months) should be ruled out. In fact, I see a fast transition as reasonably likely, for inside-view reasons that I haven't articulated and am not terribly confident about. (Further reading.) (Also relevant: Paul Christiano is well-known around here for arguing in favor of slow takeoff ... but he still assigns 30% chance of fast takeoff.)

Robin had a lot of interesting arguments in favor of slow takeoff (and long timelines, see below). He offered some inside-view arguments about the nature of intelligence and AGI, which I would counter with *different* inside-view arguments about the nature of intelligence and AGI, but that's beyond the scope of this post.

Robin also offered an outside-view argument, related to the statistics of citations in different fields—what fraction of papers get what fraction of citations? The statistics are interesting, but I don't think they shed light on the questions at issue. Take the Poincaré conjecture, which for 100 years was unproven, then all of the sudden in 2002, a reclusive genius (Perelman) announced a proof. In hindsight, we can say that

the theorem was proved gradually, with Perelman building on Hamilton's ideas from the 1980s. But really, nobody knew if Hamilton's ideas were on the right track, or how many steps away from a proof we were, until bam, a proof appeared. Likewise, no one knew how far away heavier-than-air flight was until the Wright Brothers announced that they had already done it (and indeed, people wouldn't believe them even *after* their public demonstrations). Will AGI be like that? Or will it be like Linux, developing from almost-useless to super-useful very very gradually and openly? The fact that citations are widely distributed among different papers is not incompatible with the existence of occasional sudden advances from private projects like Perelman or the Wright Brothers—indeed, these citation statistics hold in math and engineering just like everything else. The citation statistics just mean that academic fields are diverse, with lots of people working on different problems using different techniques ... which is something we already knew.

Timelines; Are we "crying wolf" about AGI?

Robin says he sees a lot of arguments that we should work on AGI prep because AGI is definitely coming soon, and that this is "crying wolf" that will discredit the field when AGI doesn't come soon. My experience is different. Pretty much all the material I read advocating for AGI safety & policy, from both inside and outside the field, is scrupulously careful to say that they do not know with confidence when we'll get AGI, and that this work is important and appropriate regardless of timelines. That doesn't mean Robin is wrong; I presume we're reading different things. I'm sure that people on the internet have said all kinds of crazy things about AGI. Oh well, what can you do?

It does seem to be an open secret that many of the people working full-time on AGI safety & policy assign a pretty high probability to AGI coming soon (say, within 10 or 20 years, or at least within their lifetimes, as opposed to centuries). I put myself in that category too. This is naturally to be expected from self-selection effects.

Again, I have inside-view reasons for privately believing that AGI has a reasonable chance of coming "soon" (as defined above), that I won't get into here. I'm not sure that this belief is especially communicable, or defensible. The party line, that "nobody knows when AGI is coming", is a lot more defensible. I am *definitely* willing to believe and defend the statement "nobody knows when AGI is coming" over an alternative statement "AGI is definitely *not* going to happen in the next 20 years". OK, well Robin didn't *exactly* say the latter statement, but he kinda gave that impression (and sorry if I'm putting words in his mouth). Anyway, I have pretty high confidence that the latter statement is unjustifiable. We even have good outside-view support for the statement "People declaring that a particular technology definitely will or won't be developed by a particular date have a terrible track-record and should be disbelieved." (see examples in There's No Fire Alarm For AGI). We don't know how many revolutionary insights lie between us and AGI, or how quickly they will come, we don't know how many lines of code need to be written (or how many ASICs need to be spun), and how long it will take to debug. We don't know any of these things. I've heard lots of prestigious domain experts talk about what steps are needed to get to AGI, and they all say different things. And they could all be wrong anyway—none of them has built an AGI! (The first viable airplane was built by the then-obscure Wright Brothers, who had better ideas than the then-most-prestigious domain experts.) Robin hasn't built an AGI either, and neither have I. Best to be humble.

Hard Problems in Cryptocurrency: Five Years Later - Buterin

This is a linkpost for https://vitalik.ca/general/2019/11/22/progress.html

Many rationalists are interested in blockchain. This article describes important mathematical problems related to blockchain, and potential solutions to cooperation problems and philanthropy via mechanism design (quadratic voting, quadratic funding).

[AN #74]: Separating beneficial AI into competence, alignment, and coping with impacts

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

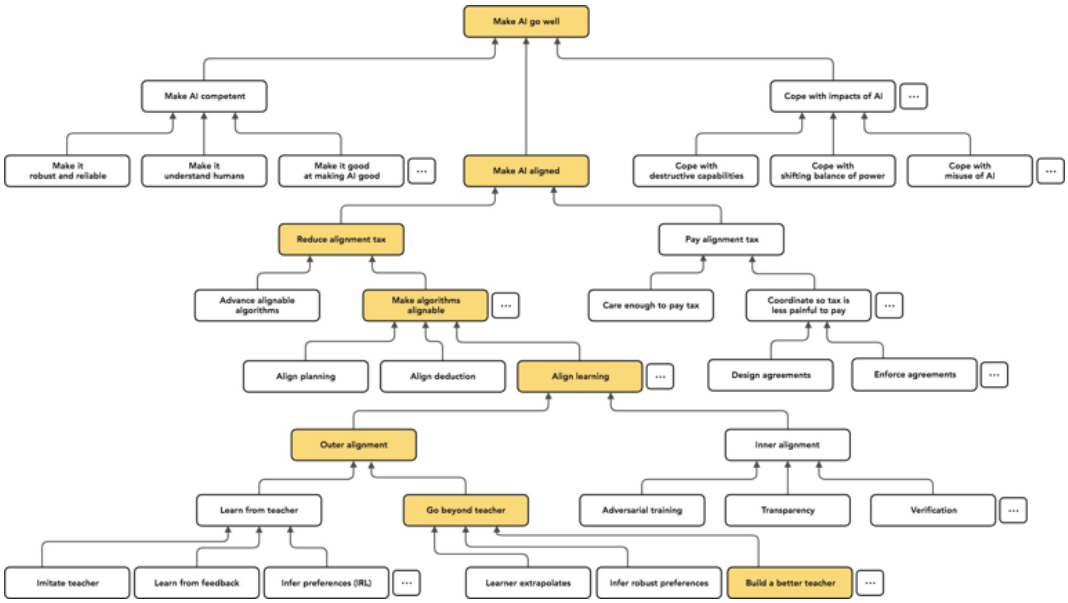
Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Highlights

AI alignment landscape *(Paul Christiano)* (summarized by Rohin): This post presents the following decomposition of how to make AI go well:

[Link to image below]



Rohin's opinion: Here are a few points about this decomposition that were particularly salient or interesting to me.

First, at the top level, the problem is decomposed into alignment, competence, and coping with the impacts of AI. The "alignment tax" (extra technical cost for safety) is only applied to alignment, and not competence. While there isn't a tax in the "coping" section, I expect that is simply due to a lack of space; I expect that extra work will be needed for this, though it may not be technical. I broadly agree with this perspective:

to me, it seems like the major technical problem which *differentially* increases long-term safety is to figure out how to get powerful AI systems that are *trying* to do what we want, i.e. they have the right motivation (AN #33). Such AI systems will hopefully make sure to check with us before taking unusual irreversible actions, making e.g. robustness and reliability less important. Note that techniques like verification, transparency, and adversarial training (AN #43) may still be needed to ensure that the *alignment* itself is robust and reliable (see the inner alignment box); the claim is just that robustness and reliability of the AI's *capabilities* is less important.

Second, strategy and policy work here is divided into two categories: improving our ability to pay technical taxes (extra work that needs to be done to make AI systems better), and improving our ability to handle impacts of AI. Often, generically improving coordination can help with both categories: for example, the publishing concerns around GPT-2 (AN #46) have allowed researchers to develop synthetic text detection (the first category) as well as to coordinate on when not to release models (the second category).

Third, the categorization is relatively agnostic to the details of the AI systems we develop -- these only show up in level 4, where Paul specifies that he is mostly thinking about aligning learning, and not planning and deduction. It's not clear to me to what extent the upper levels of the decomposition make as much sense if considering other types of AI systems: I wouldn't be surprised if I thought the decomposition was not as good for risks from e.g. powerful deductive algorithms, but it would depend on the details of how deductive algorithms become so powerful. I'd be particularly excited to see more work presenting more concrete models of powerful AGI systems, and reasoning about risks in those models, as was done in Risks from Learned Optimization (AN #58).

Previous newsletters

Addendum to AI and Compute *(Girish Sastry et al)* (summarized by Rohin): Last week, I said that this addendum suggested that we don't see the impact of AI winters in the graph of compute usage over time. While true, this was misleading: the post is measuring compute used to *train* models, which was less important in past AI research (e.g. it doesn't include Deep Blue), so it's not too surprising that we don't see the impact of AI winters.

Technical AI alignment

Mesa optimization

Will transparency help catch deception? Perhaps not *(Matthew Barnett)* (summarized by Rohin): Recent (AN #70) posts (AN #72) have been optimistic about using transparency tools to detect deceptive behavior. This post argues that we may not want to use *transparency tools*, because then the deceptive model can simply adapt to fool the transparency tools. Instead, we need something more like an end-to-end trained deception checker that's about as smart as the deceptive model, so that the deceptive model can't fool it.

Rohin's opinion: In a comment, Evan Hubinger makes a point I agree with: the transparency tools don't need to be able to detect all deception; they just need to

prevent the model from developing deception. If deception gets added slowly (i.e. the model doesn't "suddenly" become perfectly deceptive), then this can be way easier than detecting deception in arbitrary models, and could be done by tools.

Prerequisities: Relaxed adversarial training for inner alignment (AN #70)

More variations on pseudo-alignment *(Evan Hubinger)* (summarized by Nicholas): This post identifies two additional types of pseudo-alignment not mentioned in Risks from Learned Optimization (AN #58). **Corrigible pseudo-alignment** is a new subtype of corrigible alignment. In corrigible alignment, the mesa optimizer models the base objective and optimizes that. Corrigible pseudo-alignment occurs when the model of the base objective is a non-robust proxy for the true base objective. **Suboptimality deceptive alignment** is when deception would help the mesa-optimizer achieve its objective, but it does not yet realize this. This is particularly concerning because even if AI developers check for and prevent deception during training, the agent might become deceptive after it has been deployed.

Nicholas's opinion: These two variants of pseudo-alignment seem useful to keep in mind, and I am optimistic that classifying risks from mesa-optimization (and AI more generally) will make them easier to understand and address.

Preventing bad behavior

Vehicle Automation Report *(NTSB)* (summarized by Zach): Last week, the NTSB released a report on the Uber automated driving system (ADS) that hit and killed Elaine Herzberg. The pedestrian was walking across a two-lane street with a bicycle. However, the car didn't slow down before impact. Moreover, even though the environment was dark, the car was equipped with LIDAR sensors which means that the car was able to fully observe the potential for collision. The report takes a closer look at how Uber had set up their ADS and notes that in addition to not considering the possibility of jay-walkers, "...if the perception system changes the classification of a detected object, the tracking history of that object is no longer considered when generating new trajectories". Additionally, in the final few seconds leading up to the crash the vehicle engaged in *action suppression*, which is described as "a one-second period during which the ADS suppresses planned braking while the (1) system verifies the nature of the detected hazard and calculates an alternative path, or (2) vehicle operator takes control of the vehicle". The reason cited for implementing this was concerns of false alarms which could cause the vehicle to engage in unnecessary extreme maneuvers. Following the crash, Uber suspended its ADS operations and made several changes. They now use onboard safety features of the Volvo system that were previously turned off, action suppression is no longer implemented, and path predictions are held across object classification changes.

Zach's opinion: While there is a fair amount of nuance regarding the specifics of how Uber's ADS was operating it does seem as though there was a fair amount of incompetence in how the ADS was deployed. Turning off Volvo system fail-safes, not accounting for jaywalking, and trajectory reseting seem like unequivocal *mistakes*. A lot of people also seem upset that Uber was engaging in action suppression. However, given that randomly engaging in extreme maneuvering in the presence of other vehicles can *indirectly cause* accidents I have a small amount of sympathy for why such a feature existed in the first place. Of course, the feature was removed and it's worth noting that "there have been no unintended consequences—increased number of false alarms".

Read more: Jeff Kaufman writes a post summarizing both the original incident and the report. Wikipedia is also rather thorough in their reporting on the factual information. Finally, *Planning and Decision-Making for Autonomous Vehicles* gives an overview of recent trends in the field and provides good references for people interested in safety concerns.

Interpretability

Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior *(Tathagata Chakraborti et al)* (summarized by Flo): This paper reviews and discusses definitions of concepts of interpretable behaviour. The first concept, **explicability** measures how close an agent's behaviour is to the observer's expectations. An agent that takes a turn while its goal is straight ahead does not behave explicably by this definition, even if it has good reasons for its behaviour, as long as these reasons are not captured in the observer's model. **Predictable** behaviour reduces the observer's uncertainty about the agent's future behaviour. For example, an agent that is tasked to wait in a room behaves more predictably if it shuts itself off temporarily than if it paced around the room. Lastly, **legibility** or **transparency** reduces observer's uncertainty about an agent's goal. This can be achieved by preferentially taking actions that do not help with other goals. For example, an agent tasked with collecting apples can increase its legibility by actively avoiding pears, even if it could collect them without any additional costs.

These definitions do not always assume correctness of the observer's model. In particular, an agent can explicably and predictably achieve the observer's task in a specific context while actually trying to do something else. Furthermore, these properties are dynamic. If the observer's model is imperfect and evolves from observing the agent, formerly inexplicable behaviour can become explicable as the agent's plans unfold.

Flo's opinion: Conceptual clarity about these concepts seems useful for more nuanced discussions and I like the emphasis on the importance of the observer's model for interpretability. However, it seems like concepts around interpretability that are not contingent on an agent's actual behaviour (or explicit planning) would be even more important. Many state-of-the-art RL agents do not perform explicit planning, and ideally we would like to know something about their behaviour before we deploy them in novel environments.

AI strategy and policy

AI policy careers in the EU *(Lauro Langosco)*

Other progress in AI

Reinforcement learning

Superhuman AI for multiplayer poker *(Noam Brown et al)* (summarized by Matthew): In July, this paper presented the first AI that can play six-player no-limit Texas hold’em poker better than professional players. Rather than using deep learning, it works by precomputing a blueprint strategy using a novel variant of Monte Carlo linear

counterfactual regret minimization, an iterative self-play algorithm. To traverse the enormous game tree, the AI buckets moves by abstracting information in the game. During play, the AI adapts its strategy by modifying its abstractions according to how the opponents play, and by performing real-time search through the game tree. It used the equivalent of $144 of cloud compute to calculate the blueprint strategy and two server grade CPUs, which was much less hardware than what prior AI game milesones required.

Matthew's opinion: From what I understand, much of the difficulty of poker lies in being careful not to reveal information. For decades, computers have already had an upper hand in being silent, computing probabilities, and choosing unpredictable strategies, which makes me a bit surprised that this result took so long. Nonetheless, I found it interesting how little compute was required to accomplish superhuman play.

Read more: Let's Read: Superhuman AI for multiplayer poker

Meta learning

Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning *(Tianhe Yu, Deirdre Quillen, Zhanpeng He et al)* (summarized by Asya): "Meta-learning" or "learning to learn" refers to the problem of transferring insight and skills from one set of tasks to be able to quickly perform well on new tasks. For example, you might want an algorithm that trains on some set of platformer games to pick up general skills that it can use to quickly learn new platformer games.

This paper introduces a new benchmark, "Meta World", for evaluating meta-learning algorithms. The benchmark consists of 50 simulated robotic manipulation tasks that require a robot arm to do a combination of reaching, pushing and grasping. The benchmark tests the ability of algorithms to learn to do a single task well, learn one multi-task policy that trains and performs well on several tasks at once, and adapt to new tasks after training on a number of other tasks. The paper argues that unlike previous meta-learning evaluations, the task distribution in this benchmark is very broad while still having enough shared structure that meta-learning is possible.

The paper evaluates existing multi-task learning and meta-learning algorithms on this new benchmark. In meta-learning, it finds that different algorithms do better depending on how much training data they're given. In multi-task learning, it finds that the algorithm that performs best uses multiple "heads", or ends of neural networks, one for each task. It also finds that algorithms that are "off-policy"-- that estimate the value of actions other than the one that the network is currently planning to take-- perform better on multi-task learning than "on-policy" algorithms.

Asya's opinion: I really like the idea of having a standardized benchmark for evaluating meta-learning algorithms. There's a lot of room for improvement in performance on the benchmark tasks and it would be cool if this incentivized algorithm development. As with any benchmark, I worry that it is too narrow to capture all the nuances of potential algorithms; I wouldn't be surprised if some meta-learning algorithm performed poorly here but did well in some other domain.

News

CHAI 2020 Internships (summarized by Rohin): CHAI (the lab where I work) is currently accepting applications for its 2020 internship program. The deadline to apply is **Dec 15**.

How I do research

Someone asked me about this, so here are my quick thoughts.

Although I've learned a lot of math over the last year and a half, it still isn't my comparative advantage. What I do instead is,

Find a problem

that seems plausibly important to AI safety (low impact), or a phenomenon that's secretly confusing but not really explored (instrumental convergence). If you're looking for a problem, corrigibility strikes me as another thing that meets these criteria, and is still mysterious.

Think about the problem

Stare at the problem on my own, ignoring any existing thinking as much as possible. Just think about what the problem is, what's confusing about it, what a solution would look like. In retrospect, this has helped me avoid anchoring myself. Also, my prior for existing work is that it's confused and unhelpful, and I can do better by just thinking hard. I think this is pretty reasonable for a field as young as AI alignment, but I wouldn't expect this to be true at all for e.g. physics or abstract algebra. I also think this is likely to be true in any field where philosophy is required, where you need to find the right formalisms instead of working from axioms.

Therefore, when thinking about whether "responsibility for outcomes" has a simple core concept, I nearly instantly concluded it didn't, without spending a second glancing over the surely countless philosophy papers wringing their hands (yup, papers have hands) over this debate. This was the right move. I just trusted my own thinking. Lit reviews are just proxy signals of your having gained comprehension and coming to a well-considered conclusion.

Concrete examples are helpful: at first, thinking about vases in the context of impact measurement was helpful for getting a grip on low impact, even though it was secretly a red herring. I like to be concrete because we actually need *solutions* - I want to learn more about the relationship between solution specifications and the task at hand.

Make simplifying assumptions wherever possible. Assume a ridiculous amount of stuff, and then pare it down.

Don't formalize your thoughts too early - you'll just get useless mathy sludge out on the other side, the product of your confusion. Don't think for a second that having math representing your thoughts means you've necessarily made progress - for the kind of problems I'm thinking about right now, the math has to *sing* with the elegance of the philosophical insight you're formalizing.

Forget all about whether you have the license or background to come up with a solution. When I was starting out, I was too busy being fascinated by the problem to remember that I, you know, wasn't allowed to solve it.

Obviously, there are common-sense exceptions to this, mostly revolving around trying to run without any feet. It would be pretty silly to think about logical uncertainty without even knowing propositional logic. One of the advantages of immersing myself in a lot of math isn't just knowing more, but knowing what I don't know. However, I think it's rare to secretly lack the basic skills to even start on the problem at hand. You'll probably know if you are, because all your thoughts keep coming back to the same kind of confusions about a formalism, or something. Then, you look for ways to resolve the confusion (possibly by asking a question on LW or in the MIRIx Discord), find the thing, and get back to work.

Stress-test thoughts

So you've had some novel thoughts, and an insight or two, and the outlines of a solution are coming into focus. It's important not to become enamored with what you have, because it stops you from finding the truth and winning. Therefore, think about ways in which you could be wrong, situations in which the insights don't apply or in which the solution breaks. Maybe you realize the problem is a bit ill-defined, so you refactor it.

The process here is: break the solution, deeply understand why it breaks, and repeat. Don't get stuck with patches; there's a rhythm you pick up on in AI alignment, where good solutions have a certain flavor of integrity and compactness. It's OK if you don't find it right away. The key thing to keep in mind is that you aren't trying to pass the test cases, but rather to find brick after brick of insight to build a firm foundation of deep comprehension. You aren't trying to find the right equation, you're trying to find the state of mind that makes the right equation obvious. You want to understand new pieces of the world, and maybe one day, those pieces will make the difference.

ETA: I think a lot of these skills apply more broadly. Emotional trust in one's own ability to think seems important for taking actions that aren't e.g. prescribed by an authority figure. Letting myself just *think* lets me be light on my mental feet, and bold in where those feet lead me.

ETA 2: Apparently simulating drop-caps:

Like this

isn't the greatest idea. Formatting edit.

Autism And Intelligence: Much More Than You Wanted To Know

[Thanks to Marco G for proofreading and offering suggestions]

I.

Several studies have shown a genetic link between autism and intelligence; genes that contribute to autism risk also contribute to high IQ. But studies show autistic people generally have lower intelligence than neurotypical controls, often much lower. What is going on?

First, the studies. This study from UK Biobank finds a genetic correlation between genetic risk for autism and educational attainment (r = 0.34), and between autism and verbal-numerical reasoning (r = 0.19). This study of three large birth cohorts finds a correlation between genetic risk for autism and cognitive ability (beta = 0.07). This study of 45,000 Danes finds that genetic risk for autism correlates at about 0.2 with both IQ and educational attainment. These are just three randomly-selected studies; there are too many to be worth listing.

The relatives of autistic people will usually have many of the genes for autism, but not be autistic themselves. If genes for autism (without autism itself) increase intelligence, we should expect these people to be unusually smart. This is what we find; see Table 4 here. Of 11 types of psychiatric condition, only autism was associated with increased intelligence among relatives. This intelligence is shifted towards technical subjects. About 13% of autistic children have fathers who are engineers, compared to only 5% of a group of control children (though see the discussion here) for some debate over how seriously to take this; I am less sure this is accurate than most of the other statistics mentioned here).

Further (indirect) confirmation of the autism-IQ link comes from evolutionary investigations. If autism makes people less likely to reproduce, why would autism risk genes stick around in the human population? Polimanti and Gelemter (2017) find that autism risk genes aren't just sticking around. They are being *positively selected*, ie increasing with every generation, presumably because people with the genes are having more children than people without them. This means autism risk genes must be doing something good. Like everyone else, they find autism risk genes are positively correlated with years of schooling completed, college completion, and IQ. They propose that the reason evolution favors autism genes is that they generally increase intelligence.

But as mentioned before, autistic people themselves generally have very low intelligence. One study found that 69% of autistic people had an IQ below 85 (the average IQ of a high school dropout). Only 3% of autistic people were found to have IQs above 115, even though 15% of the population should be at this level.

These numbers should be taken with very many grains of salt. First, IQ tests don't do a great job of measuring autistic people. Their intelligence tends to be more imbalanced than neurotypicals', so IQ tests (which rely on an assumption that most forms of intelligence are correlated) are less applicable. Second, even if the test itself is good, autistic people may be bad at test-taking for other reasons – for example, they don't understand the directions, or they're anxious about the social interaction required to

answer an examiner's quetsions. Third, and most important, there is a strong selection bias in the samples of autistic people. Many definitions of autism center around forms of poor functioning which are correlated with low intelligence. Even if the definition is good, people who function poorly are more likely to seek out (or be coerced into) psychiatric treatment, and so are more likely to be identified. In some sense, all "autism has such-and-such characteristics" studies are studying the way people like to define autism, and tell us nothing about any underlying disease process. I talk more about this in parts 2 and 3 here.

But even adjusting for these factors, the autism – low intelligence correlation seems too strong to dismiss. For one thing, the same studies that found that relatives of autistic patients had higher IQs find that the autistic patients themselves have much lower ones. The existence of a well-defined subset of low IQ people whose relatives have higher-than-predicted IQs is a surprising finding that cuts through the measurement difficulties and suggests that this is a real phenomenon.

So what is going on here?

II.

At least part of the story is that there are at least three different causes of autism.

1. The "familial" genes mentioned above: common genes that increase IQ and that evolution positively selects for.

2. Rare "de novo mutations", ie the autistic child gets a new mutation that their non-autistic parent doesn't have. These mutations are often very bad, and are quickly selected out of the gene pool (because the people who have them don't reproduce). But "quickly selected out of the gene pool" doesn't help the individual person who got one of them, who tends to end up severely disabled. In a few cases, the parent gets the de novo mutation, but for whatever reason doesn't develop autism, and then passes it onto their child, who does develop autism.

3. Non-genetic factors. The best-studied are probably obstetric complications, eg a baby gets stuck in the birth canal and can't breathe for a long time. Pollution, infection, and trauma might also be in this basket.

These three buckets and a few other less important factors combine to determine autism risk for any individual. Combining information from a wide variety of studies, Gaugler et al estimate that about 52% of autism risk is attributable to ordinary "familial" genes, 3% to rare "de novo" mutations, 4% to complicated non-additive genetic interaction effects, and 41% "unaccounted", which may be non-genetic factors or genetic factors we don't understand and can't measure. This study finds lower heritability than the usual estimates (which are around 80% to 90%; the authors are embarrassed by this, and in a later study suggest they might just have been bad at determining who in their sample did or didn't have autism. While their exact numbers are doubtful, I think the overall finding that common familial genes are much more important than rare de novo mutations survives and is important.

Most cases of autism involve all three of these factors; that is, your overall autisticness is a combination of your familial genes, mutations, and environmental risk factors.

One way of resolving the autism-intelligence paradox is to say that familial genes for autism increase IQ, but de novo mutations and environmental insults decrease IQ.

This is common-sensically true and matches previous research into all of these factors. So the only question is whether the size of the effect is enough to fully explain the data – or whether, even after adjusting out the degree to which autism is caused by mutations and environment, it still decreases IQ.

Ronemus et al (2014) evaluate this:

They find that even autistic people without *de novo* mutations have lower-than-average IQ. But they can only screen for *de novo* mutations they know about, and it could be that they just missed some.

Here's another set of relevant graphs:

This one comes from Gardner et al (2019), which measures the cognitive ability of the fathers of autistic people and disaggregates those with and without intellectual disability. In Graph A, we see that if a child has autism (but not intellectual disability), their likelihood of having a father with any particular IQ (orange line) is *almost* the same as the likelihood of a neurotypical child having a father of that IQ (dotted line). Disguised in that "almost" is a very slight tendency for fathers to be unusually intelligent, plus a (statistically insignificant) tendency for them to be unusually unintelligent. For reasons that don't entirely make sense to me, if instead we look at the likelihood of the father to be a certain intelligence (bottom graph, where dark line surrounded by gray confidence cloud is autistic people's fathers, and dotted line is neurotypical people's fathers) it becomes more obvious that more intelligent people are actually a little more likely to have autistic children (though less intelligent people are also more likely.

(remember that "intellectual disability" just means "IQ over 70", and so many of these not-intellectually-disabled people may be very intellectually weak – I wish the paper had quantified this)

Graph B is the same thing, but with people have have autism *with* intellectual disability. Now there is a very strong effect towards their fathers being less intelligent than usual.

This confuses me a little. But for me the key point is that high-intelligence fathers show a trend (albeit not significant in this study) to be more likely than average to have children with autism *and* intellectual disability.

These questions interest me because I know a lot of people who are bright nerdy programmers married to other bright nerdy programmers, and sometimes they ask me if their children are at higher risk for autism. While their children are clearly at higher risk for autistic *traits*, I think they want to know whether they have higher risk for the most severe forms of the syndrome, including intellectual disability and poor functioning. If we take the Ronemus and Gardner studies seriously, the answer seems to be yes. The Gardner study seems to suggest it's a very weakly elevated risk, maybe only 1.1x or 1.2x relative risk. But the Gardner study also ceilings off at 90th percentile intelligence, so at this point I'm not sure what to tell these people.

III.

If Ronemus isn't missing some obscure *de novo* mutations, then people who get autism solely by accumulation of common (usually IQ-promoting) variants still end up less intelligent than average. This should be surprising; why would too many intelligence-promoting variants cause a syndrome marked by low intelligence? And how come it's so inconsistent, and many people have naturally high intelligence but aren't autistic at all?

One possibility would be something like a tower-vs-foundation model. The tower of intelligence needs to be built upon some kind of mysterious foundation. The taller the tower, the stronger the foundation has to be. If the foundation isn't strong enough for the tower, the system fails, you develop autism, and you get a collection of symptoms possibly including low intelligence. This would explain low-functioning autism from de novo mutations or obstetric trauma (the foundation is so weak that it fails no matter how short the tower is). It would explain the association of genes for intelligence with autism (holding foundation strength constant, the taller the tower, the more likely a failure). And it would also explain why there are many extremely intelligent people who don't have autism at all (you can build arbitrarily tall towers if your foundation is strong enough).

I've only found one paper that takes this model completely seriously and begins speculating on the nature of the foundation. This is Crespi 2016, Autism As A Disorder Of High Intelligence. It draws on the VPR model of intelligence, where g ("general intelligence") is divided into three subtraits, v ("verbal intelligence"), p ("perceptual intelligence"), and r ("mental rotation ability") – despite the very specific names each of these represents ability at broad categories of cognitive tasks. Crespi suggests that autism is marked by an imbalance between P (as the tower) and V + R (as the foundation). In other words, if your perceptual intelligence is much higher than your other types of intelligence, you will end up autistic.

It doesn't really present much evidence for this other than that autistic people seem to have high perceptual intelligence. Also, it doesn't really look like autistic people are worse at mental rotation. Also, the Gardner paper has analyzed autistic patients' fathers by subtype of intelligence, and there is a nonsignificant but pretty suggestive tendency for them to have higher-than-normal verbal intelligence; certainly no signs of high verbal intelligence *preventing* autism. I can't tell if this is evidence against Crespi or whether since all intellectual abilities are correlated this is just the shadow of their high perceptual intelligence, and if we directly looked at perceptual-to-verbal ratio we would see it was lower than expected. Also also, Crespi is one of those scientists who constantly has much more interesting theories than anyone else (eg), and this makes me suspicious.

Overall I would be surprised if this were the real explanation for the autism-and-intelligence paradox, but it gets an A for effort.

Conclusions

1. The genes that increase risk of autism are disproportionately also genes that increase intelligence, and vice versa (~100% confidence)

2. People diagnosed with autism are less intelligent than average (~100% confidence, leaving aside definitional complications)

3. Some of this effect is because autism is caused both by normal genes and by *de novo* mutations and environmental insults, and the *de novo* mutations and environmental insults definitely decrease intelligence. Every autism case is caused by

some combination of these three factors, and the more it is caused by normal genes, the more intelligence is likely to be preserved (~100% confidence)

4. This is not the whole story, and even cases of autism that are caused entirely or mostly by normal genetics are associated with unusually low IQ (80% confidence)

5. This can best be understood through a tower-versus-foundation model where higher intelligence that outstrips the ability of some mysterious foundation to support it will result in autism (25% confidence)

6. The specific way the model plays out may be through perceptual intelligence out of balance with verbal and rotational intelligence causing autism (3% confidence)

Is daily caffeine consumption beneficial to productivity?

Caffeine raises human alertness by binding to adenosine receptors in the human brain. It prevents those receptors from binding adenosine and suppressing activity in the central nervous system.

Regular caffeine productions seems to result in the body building more adenosine receptors, but it's unclear to me whether or not the body produces enough adenosine receptors to fully cancel out the effect. Did anybody look deeper into the issue and knows the answer?

Pricing externalities is not necessarily economically efficient

This is a linkpost for http://www.daviddfriedman.com/Academic/Coase_World.html

> [A]s long as externalities exist and are not internalized via Pigouvian taxes, the result is inefficient. The inefficiency is eliminated by charging the polluter an emission fee equal to the damage done by his pollution. In some real world cases it may be difficult to measure the amount of the damage, but, provided that that problem can be solved, using Pigouvian taxes to internalize externalities produces the efficient outcome.
>
> That analysis was accepted by virtually the entire economics profession prior to Coase's work in the field. It is wrong—not in one way but in three. The existence of externalities does not necessarily lead to an inefficient result. Pigouvian taxes, even if they can be correctly calculated, do not in general lead to the efficient result. Third, and most important, the problem is not really externalities at all—it is transaction costs.

Politics is work and work needs breaks

A post written a few years ago, posting now during a time of irrelevance (as far as I know with my limited attention on politics or social media) so as not to be accidentally taking a side in any specific political debates.

***

> **Alexandra** What has happened is shocking and everyone should oppose it.
>
> **Beatrice** I'm eating a sandwich. It is better than I expected.
>
> **Alexandra** I can't believe you would write about a sandwich at a time like this. Don't you oppose what happened?
>
> *(**Claire** is excited by the breakfast bread discussion but guesses that contributing a picture of her bagel is pretty inappropriate and looks at SMBC instead.)*

People break up their time into work and leisure. You might think of work vs. leisure as roughly 'doing stuff because other people pay for it' vs. 'doing stuff because you want to'. You can also think of it as roughly 'effortful' vs. 'relaxing'. Often these categories align—other people pay you to do effortful things, and the things you want to do are more relaxing. They don't always align. Helping your aged relative go to the doctor might be effortful but a thing you do because you want to, or your job might be so easy that you come home with a bunch of energy for challenging tasks.

I'm going to call these 'resting-' vs. 'challenged-' and '-boss' vs. '-employee'. So entirely free time is mostly resting-boss and paid work is usually challenged-employee. But you can also get resting-employee and challenged-boss activities. This all maybe relies on some models of rest and attention and effort and such that don't work out, but they seem at least close the models that most people practically rely on. For whatever reason, most people prefer to spend a decent fraction of their time doing non-effortful things, unless something terrible will happen soon if they don't.

People mostly use social media as leisure, both in the sense that nobody would intentionally pay them for it, and in the sense that it is not effortful. When important political things are happening, social media naturally turns to discussion of them. Which, if you are lucky, might be thoughtful analysis of world affairs, with a bunch of statistics and rethinking your assumptions and learning new facts about the world. Which is all great, but it is not leisure in the 'not effort' sense. When I need a break from researching the likely social consequences of artificial intelligence, moving to researching the likely social consequences of changing identity politics in America does not hit the spot as well as you might hope. I assume something similar is true of many people.

When there are opportunities to move a lot of leisure time from resting-boss idle chat to challenged-boss political discussions, people can be appalled when others do not redirect their use of social media to talking about the problem. They are thinking, 'when you are doing what you want, you should be wanting this! If you would

really spend your limited time on pictures of animals that look like pastry when you can help to stop this travesty, you are terrible!’

However this means moving time that was in ‘relaxing’ to ‘effortful’, which as far as I can tell is not super sustainable. In the sense that people usually need to spend some amount of time relaxing to be happy and able to do effortful things at other times. Redistributing all of the relaxing time to effortful time makes sense when there is a very immediate threat—for instance, your house is on fire, or you have a deadline this week that will cause you to lose your job if you don’t dedicate all of your time to it. However if you have a problem on the scale of months’ or years’ worth of effort, I think most people would favor making that effort as a sustainable trek, with breaks and lunch and jokes. For instance, if you are trying to get tenure in a few years, many would predict that you are less likely to succeed if you now attempt to ban all leisure from your life and work all of the time.

When there are political events that seem to some people to warrant talking about all of the time, and some people who really don’t want to, I think this less implies a difference in concern about the problem than you might think. The disagreeing parties could also be framing work and leisure differently, or disagreeing over how short-lived the important problem is, or when the high leverage times for responding to it are.

Self-Keeping Secrets

> A magician never reveals his secrets.

The secret behind nearly every magic trick ever performed is available at your local library. Magicial secrets stay secret because they're inconsequential. Unless you are a magician or aspire to become one, you have better things to learn than magic tricks. If magic tricks did anything that mattered then they wouldn't be magic tricks. They'd be technology.

Magicians don't need a conspiracy to keep our tricks secret. It takes work to learn how to do magic. Friction and inertia are sufficient to keep out the riffraff.

This is true of more important subjects too, like computer security. Though zero-day exploits themselves are precious secrets, "how to find" zero-days is public knowledge. And since zero-day exploits have a limited shelf-life it's "how to find" zero-days that matters.

> Three may keep a secret, if two of them are dead.
>
> —Benjamin Franklin

Organizations leak like a sponge. Organizations can keep passwords secret most of the time only because a good authentication system is easy to reset. If you're even the slightest bit concerned that your passwords have been stolen then you can re-randomize them. Similarly, an intelligence agency maintains its stockpile of zero-day exploits by constantly replenishing them. To an organization, "preserving secrecy" really means "restoring secrecy". Techniques can't be kept secret because they change too infrequently to restore secrecy after they get stolen.

In practice, organizations face the opposite problem: not enough knowledge is widely-known. Training people is so hard that the limiting factor of an organization's size is how many skilled employees it can hire. The bigger your organization gets the more it'll suffer a regression to the mean. Scaling a company is an exercise in dumbing down your employees' jobs to counteract the regression to the mean.

Large organizations can neither keep knowledge secret nor spread it around. In other words, a dependence on smart people of any kind inhibits the growth of an organization. An organization can scale to the extent it makes its employees'—and especially its customers'—intelligence unnecessary.

> SCP-055 is a "self-keeping secret" or "anti-meme".
>
> —internal document, SCP Foundation

The largest organizations are precisely those that make knowledge the most obsolete. The public school system is, by headcount, among the largest organizations in modern civilization. It must therefore, by necessity, minimize the need for students to learn anything hard[1].

Most adults are employed by large companies. Most adults buy most of our products from large companies. Small businesses are dying out[2]. Modern civilization is increasingly dominated by large organizations. These organizations don't just shape

our society. They *are* our society. We are our jobs. We are the products we use. We are the media we consume. We are our communities.

Our most popular activities are those that scale the best. Those that scale the best are those that require the least thinking, the least skill, the least specialized knowledge, the least individuality. If you want to measure your individuality, ask yourself this: of all the things you do, how much of it is so hard your friends and coworkers literally can't do it.

---

1. By "hard" I mean "conceptual". Schools can effectively force students to learn by rote. However, as a coercive institution, any school with mandatory attendance is definitionally incapable of forcing students to productively misbehave or otherwise exercise critical thinking. (Except to oppose the institution itself.) ↩

2. Small operations that concentrate a lot of talent in a tiny number of employees are doing well. These companies will continue to constitute an insignificant fraction of total employment. ↩

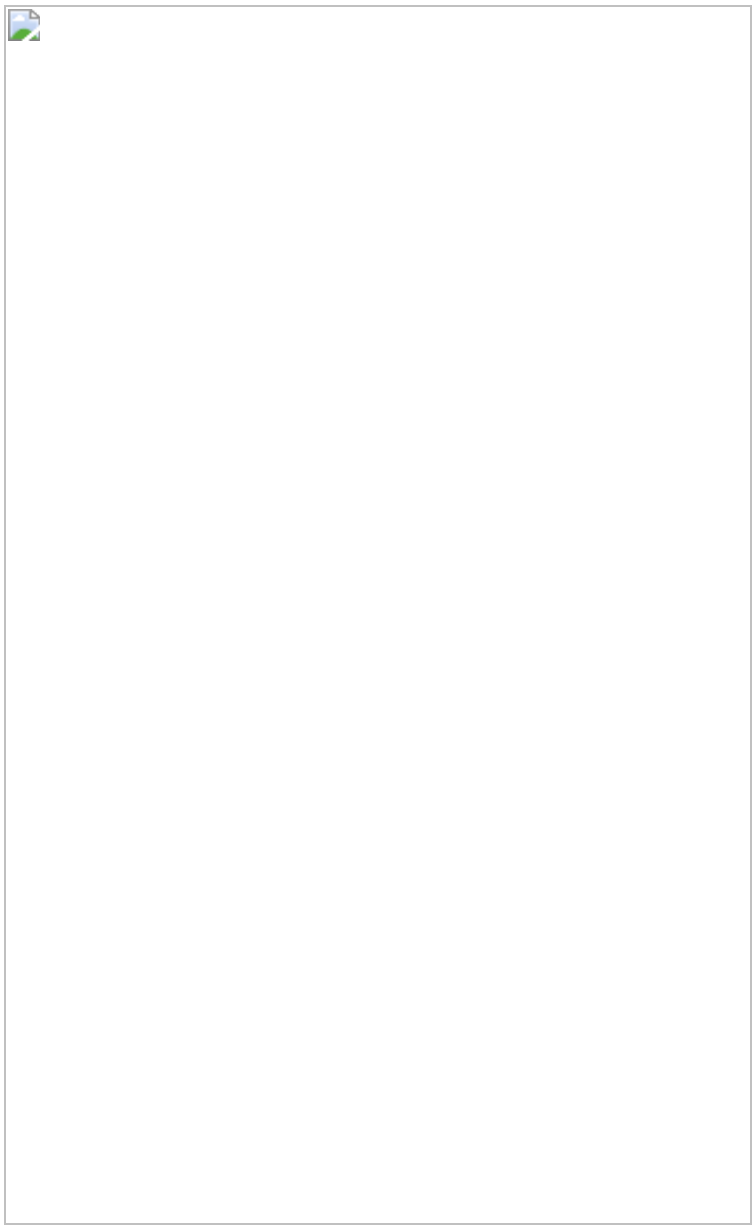
A Practical Theory of Memory Reconsolidation

Memory Reconsolidation is one candidate for a scientific of theory of "How to Actually Change Your Mind." In this post, I'll give a few fake frameworks about how memory reconsolidation works, in order to provide intuition pumps for the rest of the sequence.

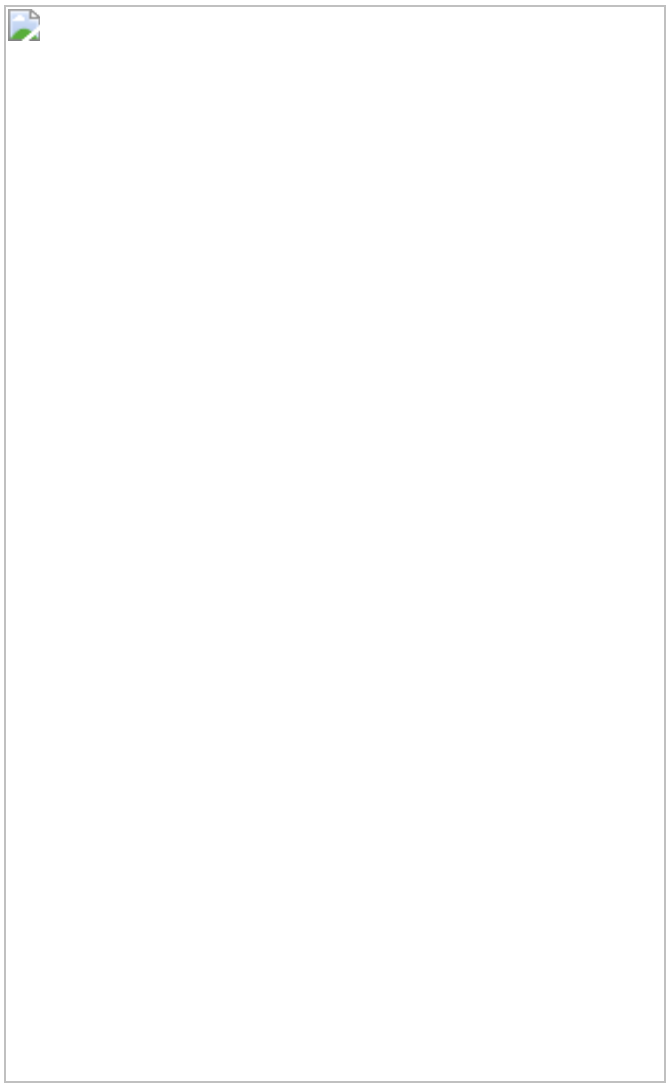
Schemas as Belief Clusters

For the purposes of this sequence, we'll describe a schema as a particular cluster of felt Aliefs that point towards a goal or set of goals.

I often find it useful to think of beliefs as arranged in a hierarchy of goals, as theorized in Perceptual Control Theory, Connection Theory, and Predictive Processing Theory.



Schemas, or Parts, then represent clusters of these beliefs trying to achieve some specific need.



In this framework, schema's don't represent any sort of natural boundary, and we can expand them and contract them at will based on what we're working on. Schemas themselves can overlap without natural boundaries, and can even contain their own internal conflicts.

3 Steps to Reconsolidate Schemas

In Unlocking the Emotional Brain, the authors state three steps needed for memory re-consolidation:

1. Activate the Old Schema

2. Challenge the Old Schema

3. Learn a Replacement Schema

All of the techniques covered in this sequence implicitly or explicitly cover the three steps. However, step #3 is also augmented later on in the Debugging Process through the "Integration" step.

The Tradeoff Between Activation and Challenge

There's an inherent tradeoff between activating the old schema and challenging it. The stronger and more perceptible the challenge is, the more likely it is to be strong enough to trigger the reconsolidation. However, this strength comes with two downsides:

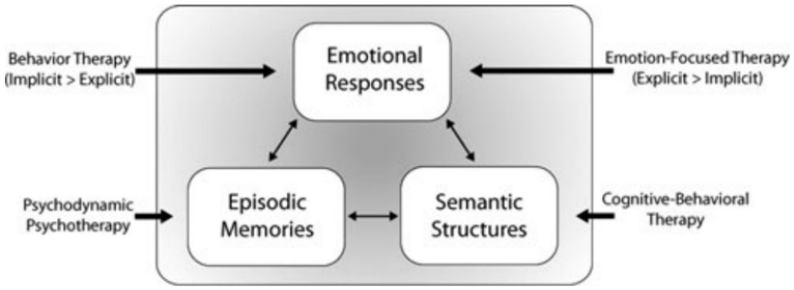
1. Challenging yourself in a very strong way doesn't feel good. It feels internally violent.

2. The stronger the challenge, the more likely you are to deactivate the schema you're challenging, creating resistance and making reconsolidation impossible.

For this reason, we start with the techniques that provide the weakest challenges, and gradually work our way up to stronger and stronger challenge, waiting till we feel a shift towards a more nuanced and accurate schema.

4 Schema Access Points

In Kaj's Post on Unlocking the Emotional Brain, he points to a paper by Lane et al that give 3 ways to access a schema:



Because I think the names for these 3 methods to access a schema are overly long and technical, I'm going to call them Felt Sense (Emotional Responses), Belief (Semantic Structures) and Evidence (Episodic Memories). Based on my own experience with changework, I'll also add a fourth category, Metaphor (metaphorical representations of a schema).

Felt Sense involves looking at the physiological or indescribable "Feelings" evoked by a schema. An example is the tightness in your throat you get when trying to express

yourself.

Belief involves the semantic content of a schema as expressed through language. An example is "If I express myself, then I will be ridiculed."

Evidence represents our internal memories of how we learned the schema. An example is blurting something out in 4th grade and being laughed at.

Metaphor represents some novel/new way of representing the schema, as an object, location, situation, or story. An example is imagining that you're wearing a mask at all times, and if you take it off people will see the silly clown face underneath and laugh at you.

Units of Action

Unit of action is a term I have been using internally to be more specific when thinking about groups of people. This post is for fleshing out and clarifying my thinking for myself, and seeing if it would be useful to anyone else. Also it feels like there really ought to be a term for this already, and I might be able to find more information if someone knows it.

Definition

A unit of action is a group that takes actions, as a group.

I take the word *unit* from the military, and also from *unit of analysis*, reflecting my belief that this is the correct level of analysis for big-picture problems. *Action* is in the colloquial sense of "did something on purpose."

To be more concrete, the kinds of things I am pointing at are like families, corporations, and government agencies. The kinds of things I am pointing away from are like race, class, gender, and religion.

Causal vs correlational

One way to see the dividing line is between groups that *cause* things to happen, and groups *to whom* things happen. Families take vacations; corporations launch products; government agencies sue people for breaking regulations. The examples I pointed away from are demographic - they don't do anything as a group, but races may be segregated, genders have different bathrooms, and classes get tax breaks.

Agent heuristic

Another heuristic is whether the group can plausibly be modeled as a single agent. Following game-theoretic intuition further, the unit of action can be broken down into other units of action the same way agents can be broken apart into multiple agents. We might model one firm as one agent when looking at the behavior of firms in competition, but when looking at the behavior of one firm model the different departments within the firm as different agents. Sometimes, as in the case of the military, these components are very explicit.

Perpetual Coordination

Units of action are stable, successful-at-some-rate coordinators. I have a vague intuition that we can think of them as coordination strategies which propagate in the same way organisms have reproduction strategies, but the analogy is hardly precise; reproduction propagates genetic information, and units of action mostly propagate *relationships.*

Miscellaneous thoughts

- Hierarchy is usually sufficient, but not necessary, to indicate a unit of action.
- There is no taxonomy; it is defined by actions, not org type.
- If a given group stops being a unit of action, I expect it to fall apart soon (split up, go bankrupt, etc).

Personal quality experimentation

Different people seem to have different strategies, which they use systematically across different parts of their lives, and that we recognize and talk about. For instance people vary on:

- Spontaneity
- Inclination toward explicit calculations
- Tendency to go meta
- Skepticism
- Optimism
- Tendency to look at the big picture vs. the details
- Expressed confidence
- Enacted patience

I don’t know of almost anyone experimenting with varying these axes, to see which setting is best for them, or even what different settings are like. Which seems like a natural thing to do in some sense, given the variation in starting positions and lack of consensus on which positions are best.

Possibly it is just very hard to change them, but my impression is that for at least some of them it is not hard to try, or to change them a bit for a short period, with some effort. (I have briefly tried making decisions faster and expressing more confidence.) And my guess is that that is enough to often be interesting. Also that if you effortfully force yourself to be more skeptical and it seems to go really well, you will find that it becomes appealing and thus easier to keep up and then get used to.

I also haven’t done this much, and it isn’t very clear to me why. Maybe it just doesn’t occur to people that much for some reason. (It also doesn’t occur to people to choose their value of time via experimentation I think, a related suggestion I like, I think from Tyler Cowen a long time ago.) So here, I suggest it. Fun date activity, maybe: randomly reselect one personality trait each, and both try to guess which one the other person is putting on.

My Anki patterns

Cross-posted from my website.

I've used Anki for ~3 years, have 37k cards and did 0.5M reviews. I have learned some useful heuristics for using it effectively. I'll borrow software engineering terminology and call heuristics for "what's good" *patterns* and heuristics for "what's bad" *antipatterns*. Cards with antipatterns are unnecessarily difficult to learn. I will first go over antipatterns I have noticed, and then share patterns I use, mostly to counteract the antipatterns. I will then throw in a grab-bag of things I've found useful to learn with Anki, and some miscellaneous tips.

Alex Vermeer's free book **Anki Essentials** helped me learn how to use Anki effectively, and I can wholeheartedly recommend it. I learned at least about the concept of interference from it, but I am likely reinventing other wheels from it.

Antipatterns

Interference

Interference occurs when trying to learn two cards together is harder than learning just one of them - one card *interferes* with learning another one. For example, when learning languages, I often confuse words which rhyme together or have a similar meaning (e.g., "vergeblich" and "erheblich" in German).

Interference is bad, because you will keep getting those cards wrong, and Anki will keep showing them to you, which is frustrating.

Ambiguity

Ambiguity occurs when the front side of a card allows multiple answers, but the back side does not list all options. For example, if the front side of a English → German card says "great", there are at least two acceptable answers: "großartig" and "gewaltig".

Ambiguity is bad, because when you review an ambiguous card and give the answer the card does not expect, you need to spend mental effort figuring out: "Do I accept my answer or do I go with Again?"

You will spend this effort every time you review the card. When you (eventually, given enough time) go with Again, Anki will treat the card as lapsed for reasons that don't track whether you are learning the facts you want to learn.

If you try to "power through" and learn ambiguous cards, you will be learning factoids that are not inherent to the material you are learning, but just accidental due to how your notes and cards represent the material. If you learn to distinguish two ambiguous cards, it will often be due to some property such as how the text is laid out. You might end up learning "great (adj.) → großartig" and "great, typeset in boldface → gewaltig", instead of the useful lesson of what actually distinguishes the words ("großartig" is

“metaphorically great” as in “what a great sandwich”, whereas “gewaltig” means “physically great” as in “the Burj Khalifa is a great structure”).

Vagueness

I carve out “vagueness” as a special case of ambiguity. Vague cards are cards where question the front side is asking is not clear. When I started using Anki, I often created cards with a trigger such as “Plato” and just slammed everything I wanted to learn about Plato on the back side: “Pupil of Socrates, Forms, wrote The Republic criticising Athenian democracy, teacher of Aristotle”.

The issue with this sort of card is that if I recall just “Plato was a pupil of Socrates and teacher of Aristotle”, I would still give the review an Again mark, because I have not recalled the remaining factoids.

Again, if you try to power through, you will have to learn “Plato → I have to recite 5 factoids”. But the fact that your card has 5 factoids on it is not knowledge of Greek philosophers.

Patterns

Noticing

The first step to removing problems is knowing that they exist and where they exist. Learn to **notice** when you got an answer wrong for the wrong reasons.

“I tried to remember for a minute and nothing came up” is a good reason. Bad reasons include the aforementioned interference, ambiguity and vagueness.

Bug tracking

When you notice a problem in your Anki deck, you are often not in the best position to immediately fix it - for example, you might be on your phone, or it might take more energy to fix it than you have at the moment. So, create a way to **track maintenance tasks** to delegate them to future you, who has more energy and can edit the deck comfortably. Make it very easy to add a maintenance task.

The way I do this is:

- I have a **big document** titled “Anki” with a structure mirroring my Anki deck hierarchy, with a list of problems for each deck. Unfortunately, adding things to a Google Doc on Android takes annoyingly many taps.
- So I also use **Google Keep**, which is more ergonomic, to store short notes marking a problem I notice. For example: “great can be großartig/gewaltig”. I move these to the doc later.
- I also use Anki’s note marking feature to note minor issues such as bad formatting of a card. I use Anki’s card browser later (with a “tag:marked” search) to fix those.

I use the same system also for tracking what information I'd like to put into Anki at some point. (This mirrors the idea from the Getting Things Done theory that *your TODO list belong outside your mind*.)

Distinguishers

Distinguishers are one way I fight interference. They are **cards that teach distinguishing interfering facts**.

For example: "erheblich" means "considerable" and "vergeblich" means "in vain". Say I notice that when given the prompt "considerable", I sometimes recall "vergeblich" instead of the right answer.

When I get the card wrong, I notice the interference, and write down "erheblich/vergeblich" into my Keep. Later, when I organize my deck on my computer, I add a "distinguisher", typically using Cloze deletion. For example, like this:

{{c1::e}}r{{c1::h}}eblich: {{c2::considerable}}

{{c1::ve}}r{{c1::g}}eblich: {{c2::in vain}}

This creates two cards: one that asks me to assign the right English meaning to the German words, and another one that shows me two English words and the common parts of the German words ("_r_eblich") and asks me to correctly fill in the blanks.

This sometimes fixes interference. When I learn the disambiguator note and later need to translate the word "considerable" into German, I might still think of the wrong word ("vergeblich") first. But now the word "vergeblich" is also a trigger for the distinguisher, so I will likely remember: "Oh, but wait, vergeblich can be confused with erheblich, and vergeblich means 'in vain', not 'considerably'". And I will more likely answer the formerly interfering card correctly.

Constraints

Constraints are useful against interference, ambiguity and vagueness.

Starting from a question such as "What's the German word for 'great'", we can add a constraint such as "… that contains the letter O", or "… that does not contain the letter E". The **constraint makes the question have only one acceptable answer** - artificially.

Because constraints are artificial, I only use them when I can't make a distinguisher. For example, when two German words are true synonyms, they cannot be distinguished based on nuances of their meaning.

In Anki, you can annotate a Cloze with a hint text. I often put the constraint into it. I use a hint of "a" to mean "word that contains the letter A", and other similar shorthands.

Other tips

Redundancy

Try to create cards using a fact in multiple ways or contexts. For example, when learning a new word, include a couple of example sentences with the word. When learning how to conjugate a verb, include both the conjugation table, and sentences with examples of each conjugated form.

Æsthethethics!

It's easier to do something if you like it. I like having all my cards follow the same style, nicely typesetting my equations with `align*`
, `\underbrace`
etc.

Clozes!

Most of my early notes were just front-back and back-front cards. Clozes are often a much better choice, because they make entering the context and expected response more natural, in situations such as:

- Fill in the missing step in this algorithm
- Complete the missing term in this equation
- Correctly conjugate this verb in this sentence
- In a line of code such as `matplotlib.pyplot.bar(x, y, color='r')`
  , you can cloze out the name of the function, its parameters, and the effect it has.

Datasets I found useful

- Shortcut keys for every program I use frequently.
  - G Suite (Docs, Sheets, Keep, etc.)
  - Google Colab
  - Vim, Vimdiff
  - Command-line programs (Git, Bash, etc.)
- Programming languages and libraries
  - Google's technologies that have an open-source counterpart
  - What's the name of a useful function
  - What are its parameters
- Unicode symbols (how to write 🐉, ←, …)
- People: first and last name ↔ photo (I am not good with names)
- English terms (spelling of "curriculum", what is "cupidity")
- NATO phonetic alphabet, for spelling things over the phone
- Mathematics (learned for fun), computer science

3 Cultural Infrastructure Ideas from MAPLE

About six months ago, I moved to the Monastic Academy in Vermont. MAPLE for short.

You may have curiosities / questions about what that is and why I moved there. But I'll save that for another time.

I was having a conversation last week about some cultural infrastructure that exists at MAPLE (that I particularly appreciate), and these ideas seemed worth writing up.

Note that MAPLE is a young place, less than a decade old in its current form. So, much of it is "experimental." These ideas aren't time-tested. But my personal experience of them has been surprisingly positive, so far.

I hope you get value out of playing with these ideas in your head or even playing with various implementations of these ideas.

1. The Care Role or Care People

MAPLE loves its roles. All residents have multiple roles in the community.

Some of them are fairly straightforward and boring. E.g. someone's role is to write down the announcements made at meals and then post them on Slack later.

Some of them are like "jobs" or "titles". E.g. someone is the bookkeeper. Someone is the Executive Director.

One special role I had the honor of holding for a few months was the Care role.

The Care role's primary aim is to watch over the health and well-being of the community as a group. This includes their physical, mental, emotional, and psychological well-being.

The Care role has a few "powers."

The Care role can offer check-ins or "Care Talks" to people. So if I, in the Care role, notice someone seems to be struggling emotionally, I can say, "Hey would you like to check in at some point today?" and then schedule such a meeting. (MAPLE has a strict schedule, and this is not something people would normally be able to do during work hours, but it's something Care can do.)

People can also request Care Talks from Care.

The Care role also has the power to plan / suggest Care Days. These are Days for community bonding and are often either for relaxation or emotional processing. Some examples of Care Days we had: we went bowling; we did a bunch of Circling; we visited a nearby waterfall.

The Care role can request changes to the schedule if they believe it would benefit the group's well-being. E.g. asking for a late wake-up. (Our usual wake-up is 4:40AM!)

Ultimately though, the point of this is that it's *someone's job* to watch over the group in this particular way. That means attending to the group field, learning how to read people even when they are silent, being attentive to individuals but also to the "group as a whole."

For me as Care, it gave me the permission and affordance to devote part of my brain function to tracking the group. Normally I would not bother devoting that much energy and attention to it because I know I wouldn't be able to do much about it even if I were tracking it.

Why devote a bunch of resource to tracking something without the corresponding ability / power to affect it?

But since it was built into the system, I got full permission to track it and then had at least some options for doing something about what I was noticing.

This was also a training opportunity for me. I wasn't perfect at the job. I felt drained sometimes. I got snippy and short sometimes. But it was all basically allowing me to train and improve at the job, as I was doing it. No one is perfect at the Care role. Some people are more suitable than others. But no one is perfect at it.

The Care role also has a Care assistant. The Care assistant is someone to pick up the slack when needed or if Care goes on vacation or something. In practice, I suspect I split doing Care Talks fairly evenly with the Care assistant, since those are a lot for one person to handle. And, people tend to feel more comfortable with certain Care people over others, so it's good to give them an option. The Care assistant is also a good person for the Care role to get support from, since it tends to be more challenging for the Care role to receive Care themselves.

I could imagine, for larger groups, having a Care Team rather than a single Care role with Care assistant.

That said, there is a benefit to having *one* person hold the mantle primarily. Which is to ensure that someone is mentally constructing a model of the group plus many of the individuals within it, keeping the bird's eye view map. This should be one of Care's main mental projects. If you try to distribute this task amongst multiple people, you'll likely end up with a patchy, stitched-together map.

In addition, understanding group dynamics and what impacts the group is another good mental project for the Care person. E.g. learning how it impacts the group when leaders exhibit stress. Learning how to use love languages to tailor care for individuals. Etc.

1.5. The Ops Role

As an addendum, it's worth mentioning the Ops role too.

At MAPLE, we follow a strict schedule and also have certain standards of behavior.

The Ops role is basically *in charge* of the schedule and the rules and the policies at MAPLE. They also give a lot of feedback to people (e.g. "please be on time"). This is a big deal. It is also probably the hardest role.

It is important for the Ops role and the Care role to not be the same person, if you can afford it.

The Ops role represents, in a way, "assertive care." The Care role represents "supportive care." These are terms about healthy, skillful parenting that I read originally from the book *Growing Up Again*.

You can read more about supportive and assertive care here.

Basically, *assertive* points to structure, and *supportive* points to nurture. Both are vital.

Care builds models of the group's physical and emotional well-being, how their interactions are going, and reading people.

Ops builds models of what parts of the structure / schedule are important, how to be fair, how to be reasonable, noticing where things are slipping, building theories as to why, and figuring out adjustments. Ops has to learn how to give and receive feedback a lot more. Ops has to make a bunch of judgment calls about what would benefit the group and what would harm the group (in the short-term and long-term), and ultimately has to do it without a higher authority telling them what to do.

It's a difficult position, but it complements the Care role very well.

As Care, I noticed that people seemed to be *worse off* and struggled more when the Ops role failed to hold a strong, predictable, and reasonable container. The Ops role is doing something that ultimately cares for people's emotional, mental, and physical well-being—same as Care. But they do it from a place of more authority and power.

As Care, I would sometimes find myself wanting to do some "Ops"-like things—like remind people about rules or structures. But it's important for Care to avoid handling those tasks, so that people feel more open and not have that "up for them" with Care. Care creates a space where people can process things and just get support.

It's not really beneficial for Care to take on the Ops role, and it's not beneficial for Ops to take on the Care role. This creates floppiness and confusion.

2. Support Committees

Sometimes, people struggle at MAPLE. Once in a while, they struggle in a way that is more consistent and persistent, in an "adaptive challenge" way. A few Care Talks aren't sufficient for helping them.

If someone starts struggling in this way, MAPLE can decide to spin up a support committee for that person. Let's call this struggling person Bob.

The specific implementation at MAPLE (as far as I know, at this particular time) is:

- Three people are selected to be on Bob's support committee.
- Some factors in deciding those people include: Is Bob comfortable with them? Do they have time? Do they want to support Bob? Do they seem like they'd do a decent job of supporting Bob?
- The way the decision actually gets made differs for each case, but it probably always involves the Executive Director.

- The support committee meets with Bob about once a week.
- They discuss ways they can be supportive to Bob. Could he use reminders to avoid caffeine? Could he use an exercise accountability person? Could he use regular Care Talks? Could he use help finding a therapist?
- They also give Bob feedback of various kinds. E.g. maybe Bob has been making chit-chat during silent periods; maybe Bob has been yelling things at Alice when he gets scared; maybe Bob is taking naps during work period. In this frame, it should be clear that Bob is the responsible party for his own growth and improvement and well-being. Ultimately he has to hold to his commitments / responsibilities / roles in the community, and the support committee can't do that *for him*. But they can help him as much as seems reasonable / worth trying.
- Current implementation of this doesn't have a pre-set deadline for when the committee ceases, but there are check-ins with the Executive Director to see how things are progressing with Bob and the support committee.
- Sometimes, it might come to make sense to ask Bob to leave the community, if things aren't improving after some time has passed (3-6 months maybe?). If everyone put in their best effort, within reason, and still Bob can't hold to his commitments, despite everyone's best intentions, then there may be a decision to part ways.
- Hopefully most of the time, the support committee thing works enough to get Bob to a place where he's no longer struggling and can get back into the flow of things without a support committee.

I appreciate support committees!

They're trying to strike a tricky balance between being supportive and holding people accountable. But they keep communication channels open and treat it like a two-way street.

Bob isn't totally in the dark about what's going on. He isn't being suddenly told there's a problem and that he can't stay. He also isn't being held totally responsible, as one might be at a normal job. "Either shape up or ship out" sort of thing. It's also not the thing where people act "open and supportive" but really it's still "on you" to fix yourself, and no one lifts a finger, and you have to do all the asking.

With a support committee, Bob gets regular support from the community in a structured way. He gets to set goals for himself, in front of others. He gets regular feedback on how he's doing on those goals. If he needs help, he has people who can brainstorm with him on how to get that help, and if they commit to helping him in some way, they actually do it. If he needs someone to talk to, he can have regularly scheduled Care Talks.

He is neither being coddled nor neglected.

It's also helpful to generally foster a feeling that the community is here for you and that there's a desire to do what's best for everyone, from all parties.

Would this kind of thing work everywhere for all groups? No, of course not.

It's a bit resource-intensive as it currently is. It also seems to ask for a high skill level and value-aligned-ness from people. But there's room to play around with the specific format.

3. The Schedule

The Schedule at MAPLE is not viable for most people in most places.

But many people who come to stay at MAPLE find out that the Schedule is something they hugely benefit from having. It's often named as one of the main pros to MAPLE life.

Basically, there's a rigid schedule in place. It applies to five-and-a-half days out of the week. (Sundays are weird because we go into town to run an event; Mondays are off-schedule days.)

But most days, it's the same routine, and everyone follows it. (The mornings and evenings are the most regimented part of the day, with more flexibility in the middle part.)

4:40AM chanting. 5:30AM meditation. 7AM exercise. 8:05AM breakfast. Then work. Etc. Etc. Up until the last thing, 8:30PM chanting.

Which is more surprising:

- The fact most people, most of the time, show up on time to each of these activities? (Where "on time" means being a little bit early?)
- Or the fact that often there's at least one person who's at least one minute late, despite there theoretically being very few other things going on, relatively speaking?

¯_(ツ)_/¯

Anyway, here's why I think the Schedule is worth talking about as a cultural infrastructure idea:

It's more conducive to getting into spontaneous motion.

You don't have to plan (as much) about what you're going to do, when. The activities come one right after the other.

At MAPLE I don't get stuck in bed, wondering whether to get up now or later.

I have spent hours and hours of my life struggling with getting out of bed (yay depression). Regardless of my mood or energy level, I just get out of bed, and it's automatic, and I don't think about it, and suddenly I'm putting on my socks, and I'm out the door.

This has translated somewhat to my off-schedule / vacation days also.

When left to my own devices, I do not exercise. I have never managed to exercise regularly as an adult. While I'm on-schedule, I just do it. I don't push myself harder than I can push; sometimes I take it easy and focus on stretching and Tai Chi. But sometimes I sprint, and sometimes I get sore, and my stamina is noticeably higher than before.

This is so much better than what it was like without the Schedule! It has proven to be more effective than my years of attempts to debug the issue using introspection.

The Schedule lets me just skip the decision paralysis. I often find myself "just spontaneously doing it." It becomes automatic habit. Like starting the drive home and "waking up" to the fact I am now home.

This is relaxing. It's more relaxing to just exercise than to internally battle over whether to exercise. It's more relaxing to just get up and start the day than to internally struggle over whether to get up. There is relief in it.

It's easier to tell when people are going through something.

As Care, it was my job to track people's overall well-being.

As it turns out, if someone starts slipping on the Schedule (showing up even a bit late to things more often), it's often an indication of something deeper.

The Schedule provides all these little fishing lines that tug when someone could use some attention, and the feedback is much faster than a verbal check-in.

Sometimes I would find myself annoyed by someone falling through or breaking policy or whatever. If I dug into it, I'd often find out they were struggling on a deeper level. Like I might find out their mom is in the hospital, or they are struggling with a flare up of chronic pain, or something like that.

Once I picked up on that pattern, I learned to view people's small transgressions or tardiness as a signal for more compassion, rather than less. Where my initial reaction might be to tense up, feel resistance, or get annoyed, I can remind myself that they're probably going through some Real Shit and that I would struggle in that situation too, and then I relax.

Everyone's doing it together.

Everyone doing something together is conducive conditions for creating common knowledge, even when there's no speaking involved. Common knowledge is a huge efficiency gain. And I suspect it's part of why it's internally easy for me to "just do it." (And maybe points to why it's harder for me to "just do it" when no one else notices or cares.)

Having more shared reality with each other reduces the need for verbal communication, formal group decision-making processes, and internal waffling.

If everyone can see the fire in the kitchen, you don't need to say a word. People will just mobilize and put out the fire.

If everyone sees that Carol is late, and Carol knows everyone has seen that she is late, it's harder for anyone to create alternative stories, like "Carol was actually on time." No one has to waste time on that.

There are lots of more flexible versions of the Schedule that people use and benefit from already. Shared meals in group houses, for instance.

But I'd love to see more experimentation with this, in communities or group houses or organizations or what-have-you.

Dragon Army attempted some things in this vein, and I saw them getting up early and exercising together on a number of occasions. I'd love to see more attempts along these lines.

Operationalizing Newcomb's Problem

The standard formulation of Newcomb's problem has always bothered me, because it seemed like a weird hypothetical designed to make people give the wrong answer. When I first saw it, my immediate response was that I would two-box, because really, I just don't believe in this "perfect predictor" Omega. And while it may be true that Newcomblike problems are the norm, most real situations are not so clear cut. It can be quite hard to demonstrate why causal decision theory is inadequate, let alone build up an intuition about it. In fact, the closest I've seen to a real-world example that made intuitive sense is Narrative Breadcrumbs vs Grizzly Bear, which still requires a fair amount of suspension of disbelief.

So, here I'd like to propose a thought experiment that would (more or less*) also work as an actual experiment.

A psychologist contacts you and asks you to sign up for an experiment in exchange for a payment. You agree to participate and sign all the forms. The psychologist tells you: "I am going to administer a polygraph (lie detector) test in which I ask whether you are going to sit in our waiting room for ten minutes after we finish the experiment. I won't tell you whether you passed, but I will give you some money in a sealed envelope, which you may open once you leave the building. If you say yes, and you pass the test, it will be $200. If you say no, or you fail the test, it will be $10. Then we are done, and you may either sit in the waiting room or leave. Please feel no obligation to stay, as the results are equally useful to us either way. The polygraph test is not perfect, but has so far been 90% accurate in predicting whether people stay or leave; 90% of the people who stay for ten minutes get $200, and 90% of those who leave immediately get $10."

You say you'll stay. You get your envelope. Do you leave the building right away, or sit in the waiting room first?

Does the answer change if you are allowed to open the envelope before deciding?

*I don't know if polygraphs are accurate enough to make this test work in the real world or not.

Building Intuitions On Non-Empirical Arguments In Science

I.

Aeon: Post-Empirical Science Is An Oxymoron And It is Dangerous:

> There is no agreed criterion to distinguish science from pseudoscience, or just plain ordinary bullshit, opening the door to all manner of metaphysics masquerading as science. This is 'post-empirical' science, where truth no longer matters, and it is potentially very dangerous.
>
> It's not difficult to find recent examples. On 8 June 2019, the front cover of New Scientist magazine boldly declared that we're 'Inside the Mirrorverse'. Its editors bid us 'Welcome to the parallel reality that's hiding in plain sight'. […]
>
> [Some physicists] claim that neutrons [are] flitting between parallel universes. They admit that the chances of proving this are 'low', or even 'zero', but it doesn't really matter. When it comes to grabbing attention, inviting that all-important click, or purchase, speculative metaphysics wins hands down.
>
> These theories are based on the notion that our Universe is not unique, that there exists a large number of other universes that somehow sit alongside or parallel to our own. For example, in the so-called Many-Worlds interpretation of quantum mechanics, there are universes containing our parallel selves, identical to us but for their different experiences of quantum physics. These theories are attractive to some few theoretical physicists and philosophers, but there is absolutely no empirical evidence for them. And, as it seems we can't ever experience these other universes, there will never be any evidence for them. As Broussard explained, these theories are sufficiently slippery to duck any kind of challenge that experimentalists might try to throw at them, and there's always someone happy to keep the idea alive.
>
> Is this really science? The answer depends on what you think society needs from science. In our post-truth age of casual lies, fake news and alternative facts, society is under extraordinary pressure from those pushing potentially dangerous antiscientific propaganda – ranging from climate-change denial to the anti-vaxxer movement to homeopathic medicines. I, for one, prefer a science that is rational and based on evidence, a science that is concerned with theories and empirical facts, a science that promotes the search for truth, no matter how transient or contingent. I prefer a science that does not readily admit theories so vague and slippery that empirical tests are either impossible or they mean absolutely nothing at all.

As always, a single quote doesn't do the argument justice, so go read the article. But I think this captures the basic argument: multiverse theories are bad, because they're untestable, and untestable science is pseudoscience.

Many great people, both philosophers of science and practicing scientists, have already discussed the problems with this point of view. But none of them lay out their argument in quite the way that makes the most sense to me. I want to do that here,

without claiming any originality or special expertise in the subject, to see if it helps convince anyone else.

II.

Consider a classic example: modern paleontology does a good job at predicting dinosaur fossils. But the creationist explanation – Satan buried fake dinosaur fossils to mislead us – also predicts the same fossils (we assume Satan is good at disguising his existence, so that the lack of other strong evidence for Satan doesn't contradict the theory). What principles help us realize that the Satan hypothesis is obviously stupid and the usual paleontological one more plausible?

One bad response: paleontology can better predict characteristics of dinosaur fossils, using arguments like "since plesiosaurs are aquatic, they will be found in areas that were underwater during the Mesozoic, but since tyrannosaurs are terrestrial, they will be found in areas that were on land", and this makes it better than the Satan hypothesis, which can only retrodict these characteristics. But this isn't quite true: since Satan is trying to fool us into believing the modern paleontology paradigm, he'll hide the fossils in ways that conform to its predictions, so we will predict plesiosaur fossils will only be found at sea – otherwise the gig would be up!

A second bad response: "The hypothesis that all our findings were planted to deceive us bleeds into conspiracy theories and touches on the problem of skepticism. These things are inherently outside the realm of science." But archaeological findings are very often deliberate hoaxes planted to deceive archaeologists, and in practice archaeologists consider and test that hypothesis the same way they consider and test every other hypothesis. Rule this out by fiat and we have to accept Piltdown Man, or at least claim that the people arguing against the veracity of Piltdown Man were doing something other than Science.

A third bad response: "Satan is supernatural and science is not allowed to consider supernatural explanations." Fine then, replace Satan with an alien. I think this is a stupid distinction – if demons really did interfere in earthly affairs, then we could investigate their actions using the same methods we use to investigate every other process. But this would take a long time to argue well, so for now let's just stick with the alien.

A fourth bad response: "There is no empirical test that distinguishes the Satan hypothesis from the paleontology hypothesis, therefore the Satan hypothesis is inherently unfalsifiable and therefore pseudoscientific." But this can't be right. After all, there's no empirical test that distinguishes the paleontology hypothesis from the Satan hypothesis! If we call one of them pseudoscience based on their inseparability, we have to call the other one pseudoscience too!

A naive Popperian (which maybe nobody really is) would have to stop here, and say that we predict dinosaur fossils will have such-and-such characteristics, but that questions like that process that drives this pattern – a long-dead ecosystem of actual dinosaurs, or the Devil planting dinosaur bones to deceive us – is a mystical question beyond the ability of Science to even conceivably solve.

I think the correct response is to say that both theories explain the data, and one cannot *empirically* test which theory is true, but the paleontology theory *is more elegant* (I am tempted to say "simpler", but that might imply I have a rigorous mathematical definition of the form of simplicity involved, which I don't). It requires fewer other weird things to be true. It involves fewer other hidden variables. It

transforms our worldview less. It gets a cleaner shave with Occam's Razor. This elegance is so important to us that it explains our vast preference for the first theory over the second.

A long tradition of philosophers of science have already written eloquently about this, summed up by Sean Carroll here:

> What makes an explanation "the best." Thomas Kuhn ,after his influential book The Structure of Scientific Revolutions led many people to think of him as a relativist when it came to scientific claims, attempted to correct this misimpression by offering a list of criteria that scientists use in practice to judge one theory better than another one: accuracy, consistency, broad scope, simplicity, and fruitfulness. "Accuracy" (fitting the data) is one of these criteria, but by no means the sole one. Any working scientist can think of cases where each of these concepts has been invoked in favor of one theory or another. But there is no unambiguous algorithm according to which we can feed in these criteria, a list of theories, and a set of data, and expect the best theory to pop out. The way in which we judge scientific theories is inescapably reflective, messy, and human. That's the reality of how science is actually done; it's a matter of judgment, not of drawing bright lines between truth and falsity or science and non-science. Fortunately, in typical cases the accumulation of evidence eventually leaves only one viable theory in the eyes of most reasonable observers.

The dinosaur hypothesis and the Satan hypothesis both fit the data, but the dinosaur hypothesis wins hands-down on simplicity. As Carroll predicts, most reasonable observers are able to converge on the same solution here, despite the philosophical complexity.

III.

I'm starting with this extreme case because its very extremity makes it easier to see the mechanism in action. But I think the same process applies to other cases that people really worry about.

Consider the riddle of the Sphinx. There's pretty good archaeological evidence supporting the consensus position that it was built by Pharaoh Khafre. But there are a few holes in that story, and a few scattered artifacts suggest it was actually built by Pharaoh Khufu; a respectable minority of archaeologists believe this. And there are a few anomalies which, if taken wildly out of context, you can use to tell a story that it was built long before Egypt existed at all, maybe by Atlantis or aliens.

So there are three competing hypotheses. All of them are consistent with current evidence (even the Atlantis one, which was written after the current evidence was found and carefully adds enough epicycles not to blatantly contradict it). Perhaps one day evidence will come to light that supports one above the others; maybe in some unexcavated tomb, a hieroglyphic tablet says "I created the Sphinx, sincerely yours, Pharaoh Khufu". But maybe this won't happen. Maybe we already have all the Sphinx-related evidence we're going to get. Maybe the information necessary to distinguish among these hypotheses has been utterly lost beyond any conceivable ability to reconstruct.

I don't want to say "No hypothesis can be tested any further, so Science is useless to us here", because then we're forced to conclude stupid things like "Science has no opinion on whether the Sphinx was built by Khafre or Atlanteans," whereas I think most scientists would actually have very strong opinions on that.

But what about the question of whether the Sphinx was built by Khafre or Khufu? This is a real open question with respectable archaeologists on both sides; what can we do about it?

I think the answer would have to be: the same thing we did with the Satan vs. paleontology question, only now it's a lot harder. We try to figure out which theory requires fewer other weird things to be true, fewer hidden variables, less transformation of our worldview – which theory works better with Occam's Razor. This is relatively easy in the Atlantis case, and hard but *potentially possible* in the Khafre vs. Khufu case.

(Bayesians can rephrase this to: given that we have a certain amount of evidence for each, can we quantify exactly how much evidence, and what our priors for each should be. It would end not with a decisive victory of one or the other, but with a probability distribution, maybe 80% chance it was Khafre, 20% chance it was Khufu)

I think this is a totally legitimate thing for Egyptologists to do, even if it never results in a particular testable claim that gets tested. If you don't think it's a legitimate thing for Egyptologists to do, I have trouble figuring out how you can justify Egyptologists rejecting the Atlantis theory.

(Again, Bayesians would start with a very low prior for Atlantis, and assess the evidence as very low, and end up with a probability distribution something like Khafre 80%, Khufu 19.999999%, Atlantis 0.000001%)

IV.

How does this relate to things like multiverse theory? Before we get there, one more hokey example:

Suppose scientists measure the mass of one particle at 32.604 units, the mass of another related particle at 204.897 units, and the mass of a third related particle at 4452.767 units. For a while, this is just how things are – it seems to be an irreducible brute fact about the universe. Then some theorist notices that if you set the mass of the first particle as x, then the second is 2πx and the third is 4/3 πx^2. They theorize that perhaps the quantum field forms some sort of extradimensional sphere, the first particle represents a diameter of a great circle of the sphere, the second the circumference of the great circle, and the third the volume of the sphere.

(please excuse the stupidity of my example, I don't know enough about physics to come up with something that isn't stupid, but I hope it will illustrate my point)

In fact, imagine that there are a hundred different particles, all with different masses, and all one hundred have masses that perfectly correspond to various mathematical properties of spheres.

Is the person who made this discovery doing Science? And should we consider their theory a useful contribution to physics?

I think the answer is clearly yes. But consider what this commits us to. Suppose the scientist came up with their Extradimensional Sphere hypothesis *after* learning the masses of the relevant particles, and so it has not predicted anything. Suppose the extradimensional sphere is outside normal space, curled up into some dimension we can't possibly access or test without a particle accelerator the size of the moon. Suppose there are no undiscovered particles in this set that can be tested to see if

they also reflect sphere-related parameters. This theory is exactly the kind of postempirical, metaphysical construct that the Aeon article savages.

But it's really compelling. We have a hundred different particles, and this theory retrodicts the properties of each of them perfectly. And it's so simple – just say the word "sphere" and the rest falls out naturally! You would have to be crazy not to think it was at least pretty plausible, or that the scientist who developed it had done some good work.

Nor do I think it seems right to say "The discovery that all of our unexplained variables perfectly match the parameters of a sphere is good, but the hypothesis that there *really is* a sphere is outside the bounds of Science." That sounds too much like saying "It's fine to say dinosaur bones have such-and-such characteristics, but we must never speculate about what kind of process produced them, or whether it involved actual dinosaurs".

V.

My understanding of the multiverse debate is that it works the same way. Scientists observe the behavior of particles, and find that a multiverse explains that behavior more simply and elegantly than not-a-multiverse.

One (doubtless exaggerated) way I've heard multiverse proponents explain their position is like this: in certain situations the math declares two contradictory answers – in the classic example, Schrodinger's cat will be both alive and dead. But when we open the box, we see only a dead cat or an alive cat, not both. Multiverse opponents say "Some unknown force steps in at the last second and destroys one of the possibility branches". Multiverse proponents say "No it doesn't, both possibility branches happen exactly the way the math says, and we end up in one of them."

Taking this exaggerated dumbed-down account as exactly right, this sounds about as hard as the dinosaurs-vs-Satan example, in terms of figuring out which is more Occam's Razor compliant. I'm sure the reality is more nuanced, but I think it can be judged by the same process. Perhaps this is the kind of reasoning that only gets us to a 90% probability there is a multiverse, rather than a 99.999999% one. But I think determining that theories have 90% probability is a reasonable scientific thing to do.

VI.

At times, the Aeon article seems to flirt with admitting that something like this is necessary:

> Such problems were judged by philosophers of science to be insurmountable, and Popper's falsifiability criterion was abandoned (though, curiously, it still lives on in the minds of many practising scientists). But rather than seek an alternative, in 1983 the philosopher Larry Laudan declared that the demarcation problem is actually intractable, and must therefore be a pseudo-problem. He argued that the real distinction is between knowledge that is reliable or unreliable, irrespective of its provenance, and claimed that terms such as 'pseudoscience' and 'unscientific' have no real meaning.

But it always jumps back from the precipice:

> So, if we can't make use of falsifiability, what do we use instead? I don't think we have any real alternative but to adopt what I might call the empirical criterion.

> Demarcation is not some kind of binary yes-or-no, right-or-wrong, black-or-white judgment. We have to admit shades of grey. Popper himself was ready to accept this, [saying]:
>
> *"The criterion of demarcation cannot be an absolutely sharp one but will itself have degrees. There will be well-testable theories, hardly testable theories, and non-testable theories. Those which are non-testable are of no interest to empirical scientists. They may be described as metaphysical."*
>
> Here, 'testability' implies only that a theory either makes contact, or holds some promise of making contact, with empirical evidence. It makes no presumptions about what we might do in light of the evidence. If the evidence verifies the theory, that's great – we celebrate and start looking for another test. If the evidence fails to support the theory, then we might ponder for a while or tinker with the auxiliary assumptions. Either way, there's a tension between the metaphysical content of the theory and the empirical data – a tension between the ideas and the facts – which prevents the metaphysics from getting completely out of hand. In this way, the metaphysics is tamed or 'naturalised', and we have something to work with. This is science.

But as we've seen, many things we really want to include as science are not testable: our credence for real dinosaurs over Satan planting fossils, our credence for Khafre building the Sphinx over Khufu or Atlanteans, or elegant patterns that explain the features of the universe like the Extradimensional-Sphere Theory.

The Aeon article is aware of Carroll's work – which, along with the paragraph quoted in Section II above, includes a lot of detailed Bayesian reasoning encompassing everything I've discussed. But the article dismisses it in a few sentences:

> Sean Carroll, a vocal advocate for the Many-Worlds interpretation, prefers abduction, or what he calls 'inference to the best explanation', which leaves us with theories that are merely 'parsimonious', a matter of judgment, and 'still might reasonably be true'. But whose judgment? In the absence of facts, what constitutes 'the best explanation'?
>
> Carroll seeks to dress his notion of inference in the cloth of respectability provided by something called Bayesian probability theory, happily overlooking its entirely subjective nature. It's a short step from here to the theorist-turned-philosopher Richard Dawid's efforts to justify the string theory programme in terms of 'theoretically confirmed theory' and 'non-empirical theory assessment'. The 'best explanation' is then based on a choice between purely metaphysical constructs, without reference to empirical evidence, based on the application of a probability theory that can be readily engineered to suit personal prejudices.

"A choice between purely metaphysical constructs, without reference to empirical evidence" sounds pretty bad, until you realize he's talking about the same reasoning we use to determine that real dinosaurs are more likely than Satan planting fossils.

I don't want to go over the exact ways in which Bayesian methods are subjective (which I think are overestimated) vs. objective. I think it's more fruitful to point out that your brain is already using Bayesian methods to interpret the photons striking your eyes into this sentence, to make snap decisions about what sense the words are used in, and to integrate them into your model of the world. If Bayesian methods are good enough to give you every single piece of evidence about the nature of the

external world that you have ever encountered in your entire life, I say they're good enough for science.

Or if you don't like that, you can use the explanation above, which barely uses the word "Bayes" at all and just describes everything in terms like "Occam's Razor" and "you wouldn't want to conclude something like that, would you?"

I know there are separate debates about whether this kind of reasoning-from-simplicity is actually good enough, when used by ordinary people, to consistently arrive at truth. Or whether it's a productive way to conduct science that will give us good new theories, or a waste of everybody's time. I sympathize with some these concerns, though I am nowhere near scientifically educated enough to have an actual opinion on the questions at play.

But I think it's important to argue that even before you describe the advantages and disadvantages of the complicated Bayesian math that lets you do this, *something like this has to be done*. The untestable is a fundamental part of science, impossible to remove. We can debate how to explain it. But denying it isn't an option.

How common is it for one entity to have a 3+ year technological lead on its nearest competitor?

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

I'm writing a follow-up to my blog post on soft takeoff and DSA, and I am looking for good examples of tech companies or academic research projects that are ~3+ years ahead of their nearest competitors in the technology(ies) they are focusing on.

Exception: I'm not that interested in projects that are pursuing some niche technology, such that no one else wants to compete with them. Also: I'm especially interested in examples that are analogous to AGI in some way, e.g. because they deal with present-day AI or because they have a feedback loop effect.

Even better would be someone with expertise on the area being able to answer the title question directly. Best of all would be some solid statistics on the matter. Thanks in advance!

[1911.08265] Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model | Arxiv

This is a linkpost for https://arxiv.org/abs/1911.08265