Becoming Stronger

1. Set Up for Success: Insights from 'Naïve Set Theory'
2. Lightness and Unease
3. The Art of the Artificial: Insights from 'Artificial Intelligence: A Modern Approach'
4. The First Rung: Insights from 'Linear Algebra Done Right'
5. Internalizing Internal Double Crux
6. Confounded No Longer: Insights from 'All of Statistics'
7. Into the Kiln: Insights from Tao's 'Analysis I'
8. Swimming Upstream: A Case Study in Instrumental Rationality
9. Making a Difference Tempore: Insights from 'Reinforcement Learning: An Introduction'
10. Turning Up the Heat: Insights from Tao's 'Analysis II'
11. And My Axiom! Insights from 'Computability and Logic'
12. Judgment Day: Insights from 'Judgment in Managerial Decision Making'
13. Continuous Improvement: Insights from 'Topology'
14. ODE to Joy: Insights from 'A First Course in Ordinary Differential Equations'
15. A Kernel of Truth: Insights from 'A Friendly Approach to Functional Analysis'
16. Problem relaxation as a tactic
17. Insights from Euclid's 'Elements'
18. Lessons I've Learned from Self-Teaching
19. Insights from Modern Principles of Economics
20. Do a cost-benefit analysis of your technology usage
21. Looking back on my alignment PhD

Set Up for Success: Insights from 'Naïve Set Theory'

Foreword

This book has been reviewed pretty thoroughly already. Rather than restate each chapter, I'll be sharing insights: some book-specific, some general.

I am quite proud of my time-to-completion for this book - just over a week, working around a very strenuous courseload. I went from having to focus really hard to pick up new concepts to reading notation nearly as fluently as the English surrounding it. The chapters started making sense - it felt less random, and more like a coherent story wherein the protagonist slowly adds Powers to their arsenal.

Naïve Set Theory

Functions

Functions $f : X \to Y$ are *just* **static sets** of ordered pairs $\{(x, f(x)) : x \in X\}$. They are *not* dynamic indexing functions, they do *not* perform efficient lookup, please do *not* waste an hour of your life trying to figure out how you could do such a thing within the simple machinery afforded by set theory up to that point.

This is one of those things that Nate talked about - how skipping over just one word a few chapters prior can cause you to waste hours. During my confusion, I knew this was probably the case, but I still couldn't manage to immediately overcome my intuitions of what a function should be. This is one reason I'm open to working through the MIRI Research Guide with others.

Families

Families are, ironically enough, just a special kind of function; don't let your intuition fool you - they aren't "groups of functions". A family belonging to $\prod_{i \in I} X_i$ maps each element i of the index set I to an element $x_i \in X_i$. For example, a family from $\{1, 2\}$ to $X_1 := \{\text{cat}, \text{dog}\}, X_2 := \{\text{tent}, \text{fire}\}$ could be $\{(1, \text{cat}), (2, \text{fire})\}$ (thanks to Dacyn for helping me clarify my writing).

Zorn's Lemma

I spent three hours staring at this proof. I understood what ZL meant. I grasped the relevant concepts. I read other versions of the proof. I still spent three long hours on this damn proof, and then I went to my classes. I don't know why I ended up figuring it out, but I suspect it was a combination of two factors: my brain worked through some things during the day, and I *really wanted it*. On the bus home, I mentally snapped and decided I was *going to understand the proof*. And I did.

I'm pleased to share my detailed proof outline of Zorn's Lemma, the product of many hours of ambient exasperation, rewritten in my own notation. Looking back, the proof in the book was pretty bad; it was neither succinct nor intuitive, but instead imposed a marais of mental variable tracking on the reader. I think mine is at least a little better, if not fully fleshed-out at all junctures.

Proof Calibration

As someone without a formal degree in mathematics, it was important for me to monitor how I approached the exercises in the book. Whenever the author began a proof, I tried to generate a mental proof sketch before reading further. Sometimes, I thought the proof would be easy and short, but it would turn out that my approach wasn’t rigorous enough. This was valuable feedback for calibration, and I intend to continue this practice. I'm still worried that down the line and in the absence of teachers, I may believe that I've learnt the research guide with the necessary rigor, go to a MIRIx workshop, and realize I hadn't been holding myself to a sufficiently high standard. Suggestions for ameliorating this would be welcome.

Forwards

Anticipation

One factor which helped me succeed was that I ensured my morning reading was what I most looked forward to each day. I was excited to go to sleep, wake up early, prepare a delicious breakfast, and curl up near the fireplace with book and paper handy. Trivial inconveniences can be fatal - do whatever you must to ensure you properly respect and anticipate your study time.

Defense with the Dark Arts

The most useful productivity-related advice I ever read was by Nate Soares (**Dark Arts warning**), and it relates to imbuing your instrumental goals with terminal values. Ever since having read that advice, every tedious assignment, every daily routine, every keystroke - they're all backed up by an intense desire to *do something* about the precarious situation in which humanity finds itself.

Internal Light

If you don't know where to start, I think the internal fire has to be lit first - don't try to force yourself to do this (or anything else) because you should. Stop the guilt-based motivation, proudly stake out what you want, and transform your life into a dazzling

assortment of activities and tasks imbued with your terminal values, your brightest visions for yourself and the future.

Lightness and Unease

Light

Month 1

January 6th is the day I finished reading *Superintelligence*, and that's the day my life felt like it entered protagonist mode. I tore through ten books over the next week and a half, from *The Art of Strategy* to *Thinking: Fast and Slow*. I proceeded to finish the considerable remainder of the Sequences I'd left unread. It wasn't fear that fueled me – it was a sweet blend of curiosity and protectiveness, about and for this cruel and beautiful world in which we live. It was a sense of the possibilities afforded to those lucky enough to live in what Scott Alexander dubbed the *carefree springtime of the universe*. It was a rekindling of my youthful inquisitiveness, when I'd spend afternoons lost in math books, coding a C++ text RPG, or caring for a caterpillar over its evolution, just to set it free.

Month 2

I applied for the upcoming CFAR workshop and was later accepted (the interview was the first time where I heard another person say things like "updating"). I continued my non-technical reading, summarized a technical alignment result, and reviewed Naïve Set Theory. The concepts in that small set theory book expanded my mathematical horizons beyond measure (okay, I haven't studied measure theory yet, so not *quite* beyond measure). This is also when things really *clicked* for me – I started generating novel insights regularly. Finally, I decided to try for 5 minutes before giving up on my CHAI application; this led to an idea of which I am truly proud.

I do not yet know whether I will be accepted for CHAI's internship, but the idea merits two posts of its own: one for the proposal itself, and one for the wonderful emotional and psychological process of discovery I experienced.

Month 3

I find myself eagerly reading chapters and completing problem sets from *AI: A Modern Approach*, well beyond what is required for my class. I'm writing more often, learning more deeply, and generally my hedonic index is through the roof relative to last fall. I know I've given the impression of having let up over the last month, but I really haven't – there are multiple projects of mine whose fruits I look forward to sharing.

Forwards

I've noticed that I go through phases of intense interest in activities, commitments, and games – this shift tended to occur multiple times per year, but has slowed as I've matured. I don't expect that to happen here. I sure hope it doesn't – I'm not ready to

relinquish this wonderful lightness, this eagerness to explore how the world *really* works on a gears level, this internal sense of purpose.

This lightness does not appear to be conditional on future acceptance or recognition – even when imagining a world in which I know with certainty that no AI safety research group brings me into their fold, the light and curiosity remain.

I am so very grateful that I found this community.

Decision

I remember the moment that was, in hindsight, a litmus test of my readiness. I was finishing the Sequences, and the reality of the "dark world" pressed down on me. My chest felt tight and pressed down upon, and gravity seemed so much *stronger* than usual; the problems were **large**, and I felt small. Would this become an obsession that would consume my psyche and my studies? Where did I leave my hero license? Who was I to make this kind of status claim, to believe that I could make progress on problems of such importance?

I came to Beyond the Reach of God and I knew I had to decide. Someone close to me had advised me to step back, to remain in the shallows for a while longer – where I was tall enough to stand. I, however, was privy to information which they were not.

Trout can swim.

Gnawing Shadows

I can sense a mix of reasonable dissatisfaction with my performance, and psychologically unrealistic expectations. I've taken far longer than I wished on my AI book; if only I were less vulnerable to pica, if I studied an extra hour each day, if the concepts had come to me more easily… I imagine worlds in which I did better and had also made substantial progress in, say, topology or linear algebra by this point in time.

I've come so far, but I also am quite aware of the countless levels above mine. Sometimes it weighs on me, but usually I view the task before me with twisted pleasure.

The Art of the Artificial: Insights from 'Artificial Intelligence: A Modern Approach'

‘ ‘ Most people won't agree to kill themselves for 50 million dollars. "

Stuart Russell and Peter Norvig

Foreword

One of the fruits of growing older is revisiting your old favorites, whether they be foods, books, or songs. As you take your warm, rose-tinted mental bath, you appreciate subtleties which had escaped you, laugh heartily at jokes which hadn't even registered, and grasp concepts whose importance you had never fully realized. Similarly, some songs never lose their charm, no matter how abused the 'replay' button becomes. *AI: AMA* is a paean to the triumphs of Computer Science and a rallying cry towards those hills we have yet to climb (*ahem*).

Exposition

My process for the first 60% of this 1,052-page behemoth was fairly normal: an hour or two of studying per day over the course of a few weeks. The last bit went a little differently.

Whatever It Takes

Two days ago, my winter trimester ended; I had upwards of 400 pages to go, half of this post to write, and dozens of exercises to complete. I found myself with three full days to spare before my research resumed the proceeding week. I did what any young man in his early twenties would do - I concocted a plan for studying upwards of 30 hours in that time span.

I knew that this plan was preposterous; in all likelihood, I'd finish 3 chapters and burn out. That's why I wheeled out Murphyjitsu.

Preparing for Failure Modes

Burnout was prepared for by:

- The imposition of some dark-arts beliefs:[1]
  - *My ego does not deplete.*
  - *I am 100% certain that I will succeed.*
  - *Given enough time, I can understand anything*.
- The institution of hourly brief exercise breaks. Cutting through the crisp, cold air at high speed got my heart beating and rekindled my passion for crushing the task at hand.
- The construction of a narrative, with me as the protagonist, slowly increasing in capability, working diligently against the racing clock.
  - Yes, I basically turned myself into a trope.
  - No, I don't mind: I'm willing to take advantage of whatever psychological factors are at my disposal to help me succeed.
  - Yes, I listened to the Rocky soundtrack a few times. It was great.

Mental exhaustion (distinct from emotional / physical burnout in that your *mind* can't process concepts as efficiently as it could seven hours ago when you started) was prepared for by:

- Regularly taking breaks to chill out, listen to music, and dance a bit. I'd also spend time dwelling on the pleasant anticipated result of having mastered the important material in the

rest of the book, making sure to *contrast it* with what would happen if I didn't follow through.
    - I made sure to avoid activities that would suck me in.
    - As outlined above - regular exercise and narrative-building.
- The pre-emptive purchase of all my favorite healthful foods; this was a great self-bribe. Having lots of delicious food on hand meant I always had something to look forward to for my breaks, and eating well meant that I didn't feel sluggish. I like to call this tactic "trivial conveniencing".
- Maintaining my regular sleep schedule of 9:30 PM - 6:00 AM.

Exercises taking forever was prevented. I performed a quick time profile of my exercise completion and found that the majority of my time was spent rereading concepts which hadn't sunk in sufficiently during the chapter. By doing relevant questions immediately after each section, I decreased time spent by about 40% and increased retention.

Wanting to do other things was mitigated by setting aside an hour or two where I'd go to the dojo or spend time with friends. I also took a bit of time for myself each night, calling my family as usual.

The Outcome

Not only did I do it, but I finished a day early. The Murphyjitsu was invaluable; the failure modes I predicted came up and were dealt with by my precautions.

24 hours of studying over the last two days, and I enjoyed every moment.

AI: A Modern Approach

If I found a concept confusing, I'll explain it in both intuitive and technical terms. Any criticisms are not directed towards the authors; it is their job to *distill* the field.

1: Introduction

In which the authors define rationality (no, IEEE Computing Edge, books do not qualify as intelligent), provide an overview of related fields, and recount a brief history of AI.

2: Intelligent Agents

In which the authors broadly define agent frameworks, environmental attributes, and the nature of learning.

Wireheading Cameo

> Notice that we said [agent performance is graded on] *environment* states, not *agent* states. If we define success in terms of agent's (sic) opinion of its own performance, an agent could achieve perfect rationality simply by deluding itself that its performance was perfect.

Of course, this division only works if there is a Cartesian boundary between that-which-grades and that-which-acts. Which there isn't.

Charming Philosophical Tangents

> The notion of "clean floor"... is based on average cleanliness over time. Yet the same average cleanliness can be achieved by two different agents, one of which does a mediocre job all the time while the other cleans energetically but takes long breaks... Which is better - a reckless life of highs and lows, or a safe but humdrum existence? Which is better - an economy where

> everyone lives in moderate poverty, or one in which some live in plenty while others are very poor? We leave these questions as an exercise for the diligent reader.

I don't know if I can answer the first question without more information, but assuming a Rawlsian veil of ignorance and considering the well-documented logarithmic hedonic effects of wealth, universal moderate poverty would be preferable. I leave the proof as an exercise for the diligent reader (it should be immediate after having read Chapter 16).

3: Solving Problems by Searching

In which the authors teach us to find what we're looking for.

Admissibility and Consistency

Heuristics are functions which estimate distance to the goal. Let g(s) be the cost to reach s in the current path, let the path cost of reaching state s' from s via action a be $c(s, a, s')$, and let h be a heuristic. The total distance function is then

$$f(s) = g(s) + h(s).$$

h is *admissible* if it never overestimates the distance remaining. Admissibility frees us from needing to exhaustively explore every path (for fear of h having hid a good solution from us through overestimation). The heuristic h is *consistent* (or, more helpfully, *monotonic*) if it obeys the triangle inequality:

$$h(s) \leq c(s, a, s') + h(s').$$

What this means: imagine that we have some admissible heuristic h_a (for example, straight-line distance on a map). An admissible but inconsistent heuristic would then be:

$$h'_a(s) = \text{hash}(s) \bmod h_a(s).$$

This is clearly admissible, but also inconsistent - every $h'_a(s)$ evaluation is some pseudo-random number between 0 and the true distance to the goal!

Claim. All consistent heuristics are admissible.

Proof. Let n_k denote a state k actions from the goal, and let d be the true distance function.

Base case (n_1):

$$\begin{aligned} h(n_1) &\leq c(n_1, a, n_0) + h(n_0) && \text{consistency} \\ &\leq c(n_1, a, n_0) && \text{definition of a heuristic} \\ &= d(n_1) && \text{definition of distance} \end{aligned}$$

Induction step ($n_k \Rightarrow n_{k+1}$):

The inductive hypothesis is that $h(n_k) \leq d(n_k)$. Then

$$\begin{aligned} h(n_{k+1}) &\leq c(n_{k+1}, a, n_k) + h(n_k) && \text{consistency} \\ &\leq c(n_{k+1}, a, n_k) + d(n_k) && \text{inductive hypothesis} \\ &= d(n_{k+1}) && \text{definition of distance .} \end{aligned}$$

Relaxation

Problem relaxation is a great way of finding admissible heuristics, and it's also a great way of approaching problems. Making potentially-unrealistic assumptions about your problem allows you to more freely explore the solution space: real-life examples include Shannon's formulation of a perfect chess algorithm in 1950, the formalization of idealized induction in Solomonoff Induction (can you induce who formalized it?), and Hutter's formulation of a perfectly rational AGI in 2000.[2]

4: Beyond Classical Search

In which the authors introduce ways to search using local information.

And-Or

Applying And-Or search can seem tricky, but it's really not. When an agent is operating under *partial observability* (it isn't sure about the exact state of the world), it maintains a *belief state* (in this chapter, the set of all states the agent could be in). To be sure it will be in the goal state after following its plan, we whip out And-Or search: **for each** state we could be in now (∧), we need to find **at least one** solution (∨).

Sometimes the environment is *nondeterministic* (actions have uncertain effects). In this case, we use And-Or search to construct a *contingency plan*, which consists of a series of if-then-else statements. Here, we control our actions, but not their effects; we will then find **at least one** action (∨) such that we have a solution **for each** potential outcome (∧). Think of this as min-maxing against an adversarial world.

5: Adversarial Search

*In which the authors demonstrate how to search when the world really **is** out to get you.*

Pruning

I won't babble about αβ-pruning - just practice the concept here. For me, it was deceptively intuitive - I "got it" so "easily" that I neglected to follow the actual algorithm in a practice problem.

Patchwork Progress

I don't think it's a good idea to spend substantial time on quick fixes which slightly improve performance but don't scale in the limit. Elegant algorithms are often superior for reasons exceeding their aesthetic appeal.

Two examples of objectively-good but non-scalable fixes:

- **Quiescence search -** sometimes, we have to stop evaluating a plan before the game is done; this can leave us in a deceptively good-looking state. Say I move my queen diagonal to your pawn and then I have to stop searching. A simple material evaluation function wouldn't see that the queen is about to get obliterated, so it returns a neutral score. Quiescence search considers the "stability" of a position and searches until it gets to quiescent states, providing a partial workaround for this problem. The search is constrained to certain types of moves.
- **Singular extension** - we try historically-good moves when we reach the search's depth limit as a last-ditch effort to extend the search tree a bit further.

This is even more relevant to deep learning, where numerous engineering tricks are employed to eke out a slightly-improved classification accuracy. I agree that spending some effort working on local optimization of established methods is beneficial, but wouldn't it be higher expected utility to have more researchers studying the fundamentals and innovating new approaches?

6: Constraint Satisfaction Problems

In which the authors show us how to color maps real pretty.

Solving n-ary CSPs

A constraint satisfaction problem (CSP) is defined as follows:

X is a set of variables, $\{X_1, \ldots, X_n\}$.

D is a set of domains, $\{D_1, \ldots, D_n\}$.

C is a set of constraints that specify allowable combinations of variables.

Sounds intimidating, but it's really not. Let's say you have two variables: $\{X_1, X_2\}$. Each can be colored red or blue, so D = $\{\{\text{red}, \text{blue}\}, \{\text{red}, \text{blue}\}\}$. Pretend that each variable is a vertex and that they're connected by an edge; we want to 2-color this graph. Then C = $\{X_1 \neq X_2\}$.

This gets tricky when we need to solve n-ary CSPs, which are those CSPs whose maximum constraint arity is n. Assume that the domains are discrete.

The main idea is that we want to break up these thick[c] constraints into mega-variables whose domains are exhaustive tuple enumerations of ways to satisfy the constraint. Then we just need to make sure that our chosen mega-solutions line up with the other mega- and normal variable assignments.

For each constraint C_k with variables S_k ($|S_k| > 2$), make a new variable y_k with domain

$$D_k' = \{ x \in \prod_{D_j \in Domains(S_k)} D_j : x \text{ satisfies } C_k \}.$$

In logical terms, D_k' is the set of all models for C_k. For each new variable, instantiate binary constraints on the identities of the values of the shared variables.

Arc consistency can be viewed in a similar light; a variable X_i is arc-consistent with another variable X_j if for each value in D_i, there exists at least once satisfactory assignment in D_j for any binary constraints between X_i, X_j. That is, every value in D_i is part of at least one model.

7: Logical Agents

In which the authors invent a formal language called "propositional logic" as a pretext for introducing us to the richest fantasy realm ever imagined: the Wumpus world.

Impish Implications

This chapter was my first time formally learning propositional logic. One thing that confused me at first: for two propositions α, β, $\alpha \Rightarrow \beta$ is true as long as it isn't the case that $\{\alpha = \text{true}, \beta = \text{false}\}$ in some model of our knowledge base. This means that even if α is *total bogus*, the implication holds.

Consider "If I live in Far Far Away, then P=NP"; α = false since I am unfortunately unable to live in that fantasy universe, and β can be either true or false - it doesn't matter here. That strange implication is logically true because *in the case where the premise is false, I make no claim about the conclusion*.

This is covered in the book, but it's important to internalize this early.

8: First-Order Logic

In which the authors generalize their newly-minted "propositional logic".

9: Inference in First-Order Logic

In which the authors incrementally introduce inference for first-order logic.

Logic Made-To-Order

When converting to conjunctive normal form, follow the steps in order. Yes, I know you *can't wait* to Skolemize, but tame your baser instincts and move your negations inwards like a good rationalist.

10: Classical Planning

In which the authors present classical planning, the child of first-order logic and search.

> Consider the noble task of buying a copy of *AI: A Modern Approach* from an online bookseller...

11: Planning and Acting in the Real World

In which the authors introduce hierarchical planning, for use in environments such the "real world" (which putatively exists).

12: Knowledge Representation

In which the authors discuss efforts to engineer ontologies and introduce modal and nonmonotonic logics.

To be frank, I didn't like the ontology-engineering portion of this chapter.

> An obvious question: do all these [special-purpose] ontologies converge on a general-purpose ontology? After centuries of philosophical and computational investigation, the answer is "Maybe".

It's possible that well-constructed ontologies are useful abstractions for agents not running on hypercomputers. However, the idea that a team of humans could engineer *one ontology to rule them all* which would produce robustly-intelligent behavior is absurd (I'm looking at the OpenMind project in particular). Set-membership ontologies consistent with reality are piecewise-linear to a nearly infinite degree, a jagged collection of edge cases and situational-truths (see: 37 Ways that Words Can Be Wrong).

13: Quantifying Uncertainty

In which the authors share a sampling of the fruits picked by Bayes and Kolmogorov, introducing the foundations of Probability Theory: prior and conditional probabilities, absolute and conditional independence, and Bayes' rule.

14: Probabilistic Reasoning

In which the authors shift their unhealthy obsession from cavities to burglaries; Bayesian networks are introduced, Markov blankets are furnished free of charge, and complimentary Monte Carlo algorithm samples are provided.

Gibbs Sampling

As a member of the Markov chain Monte Carlo algorithm family, Gibbs sampling starts from a random variable assignment (consistent with observed evidence e) and stochastically makes tweaks. The idea is to approximate the posterior distribution for whatever variable in which we are interested (the query variable).

Although the book explains this process clearly on a conceptual level, I had trouble generating the actual state transition probabilities $q(x \to x')$. Wikipedia offers more detail on the implementation than the book. It's simpler than it seems! Remember that each variable transition is governed by $P(x_i|Mb(X_i))$.

15: Probabilistic Reasoning over Time

In which the authors detail the fundamental functions of inference in temporal models and explain approaches such as hidden Markov models, the Viterbi algorithm, and particle filtering.

16: Making Simple Decisions

In which the authors introduce Probability Theory's lovely wife, Mrs. Utility Theory, and their child, Decision Theory.

Much of this chapter was familiar ground, but I found one concept counterintuitive: the non-negativity of the value of perfect information (*i.e.*, gaining information cannot decrease your expected utility). Colloquially, the value of perfect information is how much we expect to be able to improve our plans given the information. Formally,

$$VPI_e(E_j) = \left(\sum_k P(E_j = e_{jk}|e)\, EU(\alpha_{e_{jk}}|e, E_j = e_{jk})\right) - EU(\alpha|e),$$

where $EU(\alpha|e)$ is the expected utility of the best action α given the evidence e, and E_j is the random variable we just learned about.

My confusion was as follows: "suppose I believe I'm going to win a million-dollar lottery with 70% certainty, but then I get new information that I'm going to lose. Didn't my expectation decrease, no matter what action I take?". This misunderstands the VPI equation: we estimate the value of the information as a function of our *current* likelihood estimates $P(E_j = e_{jk}|e)$.

Let's shut up and multiply. Fix $P(W) = .7$, $U(W) = 1{,}000{,}000$, and $U(\neg W) = -1$.[3] Then our actions are $\{\text{buy}, \text{abstain}\}$. Under my current beliefs, buying a ticket is superior to abstaining ($EU(\text{abstain}) = 0$):

$$\begin{aligned} EU(buy) &= P(W \mid buy)U(W) + P(\neg W \mid buy)U(\neg W) \\ &= .7 \times 1{,}000{,}000 + .3 \times -1 \\ &= 699{,}999.7 \end{aligned}$$

Suddenly, a knowably-friendly oracle pops into existence and gives me the opportunity to ask one question. Being a genius, I use this question to ask *whether I will win the lottery with the next ticket I buy*. Freeze frame! Let's use theVPI equation to calculate how much this information is worth to me, *before* I receive it and under my *current* beliefs.

$$\begin{aligned} VPI(W) &= (P(W)\ EU(\alpha_{W=t} \mid W = t) + P(\neg W)\ EU(\alpha_{W=f} \mid W = f)) - EU(\alpha) \\ &= (.7 \times 1{,}000{,}000 + .3 \times 0) - 699{,}999.7 \\ &= .3 \end{aligned}$$

This advice was worth .3 utility; that is, if I knew before my purchase whether the ticket will win, 30% of the time I'd be able to avoid losing a dollar.

17: Making Complex Decisions

In which the authors show us how to make decisions in those Nashty limited-information environments.

POMDPs

Partially observable Markov decision processes model environments in which the agent can only see part of the state. Therefore, the agent performs state estimation by maintaining a probability distribution over possible physical states. It is often the case that the agent is never *quite* sure of having reached its goal; in these situations, the agent will continue to take actions to increase its belief that it has done so.

Analogously: the satisficer / maximizer dichotomy (or the *de facto* lack thereof).

18: Learning from Examples

In which the authors introduce entropy and the supervised learning techniques of yore (that is, less than a decade ago).

> [Here], we have the simplest method of all, known informally as "connect-the-dots", and superciliously as "piecewise-linear nonparametric regression".

Bayes-Structure

In supervised learning, we grade hypotheses by their likelihood given the data:

$$h^* = \underset{h \in H}{\operatorname{arg\,max}}\ P(h \mid data).$$

We apply Bayes' rule to get

$$h^* = \arg\max_{h \in H} P(data|h)P(h),$$

decomposing it into a product of goodness of fit and complexity.

✡

click

✡

19: Knowledge in Learning

In which the authors explore inductive learning, define version spaces, and introduce approaches for learning using background knowledge.

20: Learning Probabilistic Models

In which the authors introduce an assortment of acronymic algorithms and approaches, including MAP (maximum a posteriori), MLE (maximum likelihood estimation), and EM (expectation maximization).

21: Reinforcement Learning

In which the authors detail agents which learn from environmental feedback.

> It might in fact be better to learn a very simple function approximator and combine it with a certain amount of look-ahead search.

Prescient.

22: Natural Language Processing

In which the authors outline traditional approaches to text classification, information retrieval, question answering, and information extraction.

> With character models, we didn't have to worry about someone inventing a new letter of the alphabet [with the possible exception of the groundbreaking work of T. Geisel (1955)].

23: Natural Language for Communication

In which the authors outline logical and probabilistic techniques for natural language processing.

Avoiding Confusion

Let's revisit the point I made in Ch. 5 and discuss how easy it is to avoid confusion by optimizing based on what you know how to do *now* - this seems to be a common and serious failure mode. Half of this chapter is about efforts to contort English to fit inside hard-and-fast syntactic and semantic rules (which are either provided or learned).

Imagine that your research involves explaining and predicting the shape taken by water in various situations. You decide to define a probability distribution over "shapes that water can

take" and a transition model for "how the shapes change over time". You might start with easy examples (such as vase- and bucket-"shaped" water), leaving the hard problems (like ocean- and water vapor-"shaped" water) to future researchers. You "solve" the easy cases and get a little more ambitious.

With the help of a team of professional fluidicians, you enumerate "common-sense" rules such as:

> In simplified systems consisting of raindrops and the ground, the shape of each discrete body of water can be represented by a conditional multivariate Gaussian, where the distribution is conditional on whether it has struck the ground yet.

You set up high-FPS cameras in storms and collect video data for millions of raindrop-impact events. You're even able to avoid manual processing via MTurk by employing the latest advances in deep learning! You use segmentation to automatically isolate raindrop pixels and a pretrained recurrent network to detect the frame of impact, allowing for easy classification of all other frames as pre-impact or post-impact. Since you read Ch. 20, you know you can use maximum-likelihood estimation to learn the parameters for your conditional multivariate Gaussian using your newly-labelled water shapes.

But what if a raindrop strikes a sharp corner, splaying the drop's water in many directions? Obviously, you just need another edge case - a StruckSharpCorner condition in the Gaussian. For that, you go to gather more data...

Or you could derive fluid dynamics.[4]

24: Perception

In which the authors illuminate low-level details of our visual system and outline how these insights have been applied to computer vision tasks.

25: Robotics

In which the authors combine myriad methods from their opus to tackle the difficult problem of robotic control.

Holonomy

"Holonomic" and "nonholonomic" were words I never knew I wanted so badly (similar to "ontology" and "epistemology").

Technically, a robot is holonomic if

$$|\text{ effective degrees of freedom }| \;=\; |\text{ controllable degrees of freedom }|\,.$$

Imagine you're driving a car on a Cartesian plane. Your car can reach any (x, y) point and end up in any orientation θ you so choose, giving it three effective degrees of freedom (even though you can only turn and drive forwards / backwards). Cars are then nonholonomic, since $3 \neq 2$ (the proof of which is left as an exercise to the dedicated reader). A car which could also move *sideways* would be holonomic.

Alignment, Solved

I bring ye good tidings! Russell and Norvig introduce a full solution to the control problem: the alignment method.

> The object is represented by M features or distinguished points m_1, m_2, . . . , m_M in three-dimensional space...

Oh.

26: Philosophical Foundations

In which the authors consider a range of ethical and philosophical quandries in AI.

Underestimating Books in the Chinese Room

> The rule book and the stacks of paper, being just pieces of paper, do not understand Chinese.

John Searle obviously doesn't read IEEE Computing Edge.

No Universal Arguments

> One can hope that a robot that is smart enough to figure out how to terminate the human race is also smart enough to figure out that that was not the intended utility function.

Why would it care?

27: AI: The Present and Future

In which the authors introduce one last concept, asymptotic bounded optimality, and look forward to the great tasks ahead.

> ‘ ‘ We can see only a short distance ahead, but we can see that much remains to be done. "
>
> Alan Turing

Final Thoughts

The authors wield light-hearted prose regularly and to great effect; I often found myself chuckling heartily. Although the pages are packed and the book is big, fear not: if you pay attention and become invested in the task at hand, reading *AI: AMA* constitutes quite the enjoyable journey.

This seems like a good book to read early on in the MIRI reading list.

Tips

- Do exercises immediately after each section in the chapter.
  - I randomly chose ~7 exercises per chapter (excluding the programming problems), only skipping exercises which looked trivial or not relevant.
- Chegg was the only reputable place I could find with an answer key, although the answers were often of low quality. I'd recommend just using StackExchange.
- Feel free to skip exercises for the following chapters: 1, 11, 12, 19, 22-27.

Forwards

Meta

Writing

For the first half of the book, I didn't write each chapter's commentary immediately after reading. This was a mistake. Skimming half of a thousand-page book to remember what I got stuck on is **not** my idea of a good time.

Conceptual Issues

Proofs remain inordinately difficult for me, although I have noticed a small improvement. To do MIRI-relevant math, proofs will need to become second nature. Depending on how I feel as I progress through my next book (which will likely be a proof-centric linear algebra tome), I'll start trying different supplemental approaches for improving my proof prowess.

I have resolved that by the completion of my next book review, proofs will be one of my strong suits.

Theoretical machine learning is another area in which I still notice pangs of confusion. I didn't stop to work on this too much, as I plan to revisit formal machine learning after developing more mathematical sophistication. I'm comfortable with deep learning and its broad-strokes conceptual backdrop, but I don't like not having my gears-level comprehension in order.

Study Group

If you're interested in working through this book (or other books on the reading list) with me or others, there is a MIRIx Discord run by Diffractor. For an invite link, feel free to message me!

Subjective

Anticipation

I'll admit it: while reading, there were many times when my heart began to race. I was thrilled, realizing that I was *finally going to learn* a concept about which I had been curious for oh-so-long. For example, I had worked with the expectation maximization algorithm early in my undergraduate years; at the time, it was a Sophisticated-Person Concept, well beyond my reach.

And then I learned the concept, just like a normal person. It wasn't even particularly hard.

Benefits

I've noticed that learning this content has not only advanced me towards my MIRI-centric goals, but also improved my ability to excel in both my classes and my research (which is on computer-aided molecule generation). I suspect this trend will continue.

As I've worked to increase my scholarship and understanding of the (computational) world, I've become more and more aware of exactly how much I do not know. This excites me. I can't wait to learn Information Theory, Statistics, and Topology, to name a few.

I feel a bit like a kid in a candy shop.

On "Difficulty"

I am convinced that *there are no hard concepts*, only concepts which take different amounts of time to learn.[5] This is not trivial; dissolving the seemingly ontologically-basic "difficult for me"

attribute goes a long way towards having the persistence to figure things out.

Given enough time, I can understand anything.

---

[1] From personal experience, I don't recommend using this technique liberally; it's already hard enough to correct our epistemologies.

[2] Disclaimer: even setting aside the need for a hypercomputer, AIXI has issues (such as not being naturalized). This isn't the droid you're looking for.

[3] Assume utility scales linearly with money for simplicity; for similar reasons, I'm blending evidence and state variables. Shame on me!

[4] Artificial Addition talks about this confusion avoidance.

[5] Wording credited to Diffractor.

The First Rung: Insights from 'Linear Algebra Done Right'

Foreword

Linear algebra, my old flame - how I missed you. At my undergraduate institution, linear algebra was my introduction to proof-based mathematics. There are people who shake hands, and there are people who **shake hands**. You know the type - you grasp their hand, and they clamp down and pull you in, agitating so wildly you fear for the structural integrity of your joints. My first experience with proofs was an encounter of the latter variety.

I received my first homework grade, and I was *not* pleased with my performance. I promptly went to the library and vowed not to leave until I learned how to write proofs adequately. The hours passed, and, (thankfully for my stomach), I got it. I didn't let up all semester. Immediately before the final exam, I was doing pushups in the hallway, high-fiving my friends, and watching the 'Michael Jordan Top 50 All Time Plays' video while visualizing myself doing that to the test. Do that to the test I did indeed.

This time around, the appropriately-acronymized *LADR* is the first step on my journey to attain a professional-grade mathematical skillset.

Tight Feedback Loops

In a (possibly maniacal) effort to ensure both mastery of the material and the maturation of my proof skillset, I did nearly[1] every one of the 561 exercises provided. I skipped problems only when I was confident I wouldn't learn anything, or calculus I didn't remember was required (and the payoff didn't seem worth the time spent relearning it now in a shallow manner, as opposed to thoroughly learning more calculus later). If I could sketch a solid proof in my head, I wouldn't write anything down. Even in the latter case, I checked my answers using this site (additional solutions may be found here, although be warned that not all of them are correct).

I also sometimes elected to give myself small hints after being stuck on a problem for a while; the idea was to keep things at the difficulty sweet spot. Specifically, I'd spend 10-20 minutes working on a problem by myself; if I wasn't getting anywhere, I'd find a hint and then *backpropagate the correct mental motion instead of what I had been trying to do.* I think that focusing on where you were going wrong and what insight you *should* have had, in what direction you *should* have looked, is more efficient than just reading solutions.

Over time, I needed fewer hints, even as problem difficulty increased.

My approach was in part motivated by the findings of Rohrer and Pashler:

> Surprisingly little is known about how long-term retention is most efficiently achieved... Our results suggest that a single session devoted to the study of some

> material should continue long enough to ensure that mastery is achieved but that immediate further study of the same material is an inefficient use of time.

The point isn't to struggle *per se* - it's to improve and to *win*.

L**inear** A**lgebra** D**one** R**ight**

This book has been previously reviewed by Nate Soares; as such, I'll spend time focusing on the concepts I found most difficult. Note that his review was for the second edition, while mine is for the third.

True to my vow in the last post, I have greatly improved my proof-babble; a sampling of my proofs can be found here.

> If you zip through a page in less than an hour, you are probably going too fast.

Try me.

1: Vector Spaces

In which the author reviews complex numbers, vector spaces, and subspaces.

I kept having trouble parsing

> For $f, g \in F^S$, the sum $f + g \in F^S$ is defined by $(f + g)(x) = f(x) + g(x)$ for all $x \in S$.

because my brain was insisting there was a type error in the function composition. I then had the stunning (and overdue) realization that my mental buckets for "set-theoretic functions" and "mathematical functions in general" should be merged.

That is, if you define

$$f : X \to Y = \{ (x, f(x)) : x \in X \}$$

$$g : X \to Y = \{ (x, g(x)) : x \in X \},$$

then $(f + g) : X \to Y$ simply has the definition $\{(x, f(x) + g(x)) : x \in X\}$. There isn't "online computation"; the composite function simply has a different Platonic lookup table.

2: Finite-Dimensional Vector Spaces

In which the author covers topics spanning linear independence, bases, and dimension.

3: Linear Maps

In which the author guides us through the fertile territory of linear maps, introducing null spaces, matrices, isomorphisms, product and quotient spaces, and dual bases.

> So far our attention has focused on vector spaces. No one gets excited about vector spaces.

Matrix Redpilling

The author built up to matrix multiplication by repeatedly insinuating that linear maps are secretly just matrix multiplications, teaching you to see the true fabric of the territory you've been exploring. Very well done.

Look no further than here and here for an intuitive understanding of matrix multiplication.

Dual Maps

> If $T \in L(V, W)$ then the dual map of T is the linear map $T' \in L(W', V')$ defined by $T'(\phi) = \phi \circ T$ for $\phi \in W'$.

This StackExchange post both articulates and answers my initial confusion.

Grueling Dualing

> The double dual space of V, denoted V'', is defined to be the dual space of V'. In other words, $V'' = (V')'$. Define $\Lambda : V \to V''$ by $(\Lambda v)(\varphi) = \varphi(v)$ for $v \in V$ and $\varphi \in V'$.

Stay with me, this is dualble.

So Λ takes some $v \in V$ and returns the curried function $\Lambda_v \in V''$. Λ_v, being in V'', takes some $\varphi \in V'$ and returns some $a \in F$. In other words, $\Lambda_v \in V''$ lets you evaluate the space of evaluation functions (V') with respect to the *fixed* $v \in V$. That's it!

4: Polynomials

In which the author demystifies the quadratic formula, sharing with the reader those reagents used in its incantation.

> Remarkably, mathematicians have proved that no formula exists for the zeros of polynomials of degree 5 or higher. But computers and calculators can use clever numerical methods to find good approximations to the zeros of any polynomial, even when exact zeros cannot be found.

> For example, no one will ever be able to give an exact formula for a zero of the polynomial p defined by $p(x) = x^5 - 5x^4 - 6x^3 + 17x^2 + 4x - 7$.

...

There are two cats where I live. Sometimes, I watch them meander around; it's fascinating to think how they go about their lives totally oblivious to the true nature of the world around them. The above incomputability result surprised me so much that I have begun to suspect that I too am a clueless cat (until I learn complex analysis; you'll excuse me for having a few other textbooks to study first).

Edit: daozaich writes about why this isn't as surprising as it seems.

5: Eigenvalues, Eigenvectors, and Invariant Subspaces

*In which the author uses the prefix '*eigen*-' so much that it stops sounding like a word.*

Revisiting Material

Before starting this book, I watched 3Blue1Brown's video on eigenvectors and came out with a vague "understanding". Rewatching it after reading Ch. 5.A, the geometric intuitions behind eigenvectors didn't seem like useful ways-to-remember an exotic math concept, they felt like a manifestation of how the world works. I *knew* what I was seeing from the hundreds of proofs I'd done up to that point.

Imagine being blind yet knowing the minute details of each object in your room; one day, a miracle treatment restores your eyesight in full. Imagine then seeing your room for the "first time".

Diagonalizability

Intuitively, the diagonalizability of some operator $T \in L(V)$ on a finite-dimensional vector space V means you can partition (more precisely, express as a direct sum) V by the eigenspaces $E(\lambda_i, T)$.

Another way to look at it is that diagonalization is the mutation of the basis vectors of V so that each column of $M(T)$ is one-hot[2]; you then rearrange the columns (by relabeling the basis vectors) so that $M(T)$ is diagonal.

Unclear Exercise

On page 156, you'll be asked to verify that a matrix is diagonalizable with respect to a provided nonstandard basis. The phrasing of the exercise makes it seem trivial, but

the book doesn't specify how to do this until Ch. 10. Furthermore, it isn't core conceptual material. Skip.

6: Inner Product Spaces

In which the author introduces inner products, orthonormal bases, the Cauchy-Schwarz inequality, and a neat solution to minimization problems using orthogonal complements.

7: Operators on Inner Product Spaces

In which the author lays out adjoint, self-adjoint, normal, and isometric operators, proves the (a) Spectral theorem, and blows my mind with the Polar and Singular Value Decompositions.

Adjoints

Consider the linear functional $\varphi \in L(W, F)$ given by $\langle Tv, w \rangle$ for fixed $v \in V$; this is then a linear functional on W for the chosen Tv. The adjoint T^* produces the corresponding linear functional in $L(V, F)$; given fixed $w \in W$, we now map to some linear functional on V such that $\langle Tv, w \rangle = \langle v, T^*w \rangle$. The left-hand side is a linear functional on W, and the right-hand side is a linear functional on V.

The Ghost Theorem

My brain was unreasonably excited for this chapter because I'd get to learn about "ghosts" (AKA the Spectral theorem). My conscious self-assurances to the contrary completely failed to dampen this ambient anticipation.

8: Operators on Complex Vector Spaces

In which generalized eigenvectors, nilpotent operators, characteristic and minimal polynomials, and the Jordan Form make an appearance, among others.

9: Operators on Real Vector Spaces

In which real vector spaces are complexified and real operators are brought up to speed with their complex counterparts.

10: Trace and Determinant

In which the curtain is finally pulled back.

> We proved the basic results of linear algebra before introducing determinants in this final chapter. Although determinants have value as a research tool in more advanced subjects, they play little role in basic linear algebra (when the subject is **done right**).

Sassy partial title drop (emphasis mine).

Final Verdict

Overall, I really liked this book and its clean theoretical approach. By withholding trace and det until the end of the book, many properties were arrived at in a natural, satisfying, and enlightening manner. The proofs were clean, and the writing was succinct (although I did miss the subtle wit of Russell and Norvig). This book positively, definitely belongs on the MIRI book list.

Forwards

Timing

This review follows the previous by exactly four weeks; however, I was at CFAR for a week during that time, dedicated a few days to *All of Statistics* (my next review), and slowed myself considerably by doing **five hundred** proofs. If I were treating this as a normal textbook, I imagine it would have taken less than half the time.

The most exciting effect of diving into math like this is that when I don't understand a concept, I now eagerly look to the *formalization* for clarification (previously, I'd barely be able to track all the Greek).

Fluency

An interesting parallel between learning math and learning languages: when I started picking up French, at first the experience was basically always was "ugh now I have to look up 5 things to even have the vocabulary to ask how to turn on my computer". Eventually, it became natural to belt out *et comment est-ce que je peux allumer mon ordi ? L'enfoiré refuse de fonctionner, comme d'habitude ; c'est grand temps que j'en achète un de plus*. No checking needed.

And so it went with proofs - "what techniques can I use to translate this statement into the answer" turned into "the proof feels like it's flowing out of my arm?!".

Proofs

I've noticed that when I successfully produce a non-trivial proof, it's nearly always when I have a strong understanding that *this is how the world is*. The proof is then just translating this understanding to math-ese, pounding away at the shell of the problem with every tool at my disposal to reach this truth.

Imagine a friend of yours fell under the ice. In one situation, you meander, blindfolded and half-deaf, with a vague idea of "I *think* they were this way?", trying different things and occasionally hearing faint pounding.

Now consider the situation in which you *know* where they are; it's then a matter of finding the right tools to smash the ice. You strike with everything you have, with every ounce of strength you possess; finally, you break your friend free.

Impatience

My most obvious remaining weak point with proofs is impatience. I have a strong intuition that this impulse is borne from my programming experience. When I write code, I carefully consider pre- and post-conditions, expected use cases, and the context of the problem. When using an external library, things are different; when asked why something is appropriate for use in a (low-stakes) program, it's fine to only provide high-level intuitions.

Similarly, in the few situations in which I have had to prove a novel result, I have found myself being extremely cautious (and rightly so). However, when proving a known result, a strong desire to take shortcuts overtakes me. I'm going to have to keep ironing this out.

Hiding Ignorance

Another aspect of this journey which I greatly enjoy is the methodical elimination of deficiencies and weak points. In my deep learning class, I had great trouble remembering what an eigenvalue was - it was at this moment that I knew I had to get down to business. Working with a surface-level understanding yields superficial results.

I imagine I was not the only person who was somewhat confused. However, being the first to admit confusion feels low-status: "everyone else seems to be following along, so I better be quiet and figure this out on my own time." I've made a point of ignoring this reasoning and asking more questions, and I think it's paid off. Incidentally, everyone else seemed relieved that the question got asked.

LessWrong

I'd like to add that in these posts, I present a somewhat distorted perspective of my academic life; these weak points are the exception, not the norm (*ahem*). I focus on my weak points because I want to become stronger - to admit them is not necessarily to say "*I* am weak" (although this may be the case relative to the person I want to become).

Speaking from experience, I feel that this is intimidating to newcomers. The culture can appear highly critical; this has been discussed before. I hope to do my part through these very posts, in which I plainly admit "I forgot eigenvalues. I fixed it - and I'm better off for having done so."

Calculus

The calculus-based exercises in this book and in *All of Statistics* make me uncomfortable. In the spirit of not hiding ignorance, I'll admit it[3] - I totally forgot how to integrate by parts, among other things. Although MIRI math is mostly discrete, I imagine that I'll still make a quick run through a calculus textbook in the near future.

I also find myself curious about real and complex analysis, but I suspect that's more of a luxury (given timelines). Maybe I'll learn it in my free time at some point.

Lost Calling

I have the distinct feeling of having been incredibly silly for many years; one of the reasons being my pretending that I didn't love math. In high school, I did quite well (and was designated the outstanding mathematics student of my class) as a product of my passion for toying with math in my free time.

However, in college, I just wanted to learn computer science. I'd gloss over the low-level math (although I did do some Project Euler for fun). Instead, I preferred learning to find clever high-level solutions and build up an algorithm-centric problem-solving toolkit. Now that I've truly taken the plunge, the water is just so nice.

I'm sorry to have been away for so long.

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also have the pleasant side effect of your receiving an invitation to the MIRIx Discord server.

[1] For Ch. 8-10, I did a random sampling of 15% of the practice problems, as opposed to 100% (I was reaching steeply-diminishing returns for relevant gears-level gains).

[2] Please let me know if there's a more appropriate linear-algebraic term for this.

[3] Merely admitting ignorance is not virtuous.

> The eighth virtue is humility. To be humble is to take specific actions in anticipation of your own errors. To confess your fallibility and then do nothing about it is not humble; it is boasting of your modesty. Who are most humble? Those who most skillfully prepare for the deepest and most catastrophic errors in their own beliefs and plans. Because this world contains many whose grasp of rationality is abysmal, beginning students of rationality win arguments and acquire an exaggerated view of their own abilities. But it is useless to be superior: Life is not graded on a curve. The best physicist in ancient Greece could not calculate the path of a falling apple. There is no guarantee that adequacy is possible given your hardest effort; therefore spare no thought for whether others are doing worse. If you compare yourself to others you will not see the biases that all humans share. To be human is to make ten thousand errors. No one in this world achieves perfection.

The virtue is in shedding ignorance:

The first virtue is curiosity. A burning itch to know is higher than a solemn vow to pursue truth. To feel the burning itch of curiosity requires both that you be ignorant, and that you desire to relinquish your ignorance. If in your heart you believe you already know, or if in your heart you do not wish to know, then your questioning will be purposeless and your skills without direction. Curiosity seeks to annihilate itself; there is no curiosity that does not want an answer. The glory of glorious mystery is to be solved, after which it ceases to be mystery. Be wary of those who speak of being open-minded and modestly confess their ignorance. There is a time to confess your ignorance and a time to relinquish your ignorance.

~ *The Twelve Virtues of Rationality*

Internalizing Internal Double Crux

> In sciences such as psychology and sociology, internalization involves the integration of attitudes, values, standards and the opinions of others into one's own identity or sense of self.

Internal Double Crux is one of the most important skills I've ever learned. In the last two weeks, I've solved some serious, long-standing problems with IDC (permanently, as far as I can tell, and often in less than *5 minutes*), a small sample of which includes:

- Belief that I have intrinsically less worth than others
- Belief that others are intrinsically less likely to want to talk to me
- Belief that attendance at events I host is directly tied to my worth
- Disproportionately negative reaction to being stood up
- Long-standing phobia of bees and flies

I feel great, and I love it. Actually, most of the time I don't feel amazingly confident - I just feel *not bad* in lots of situations. Apparently this level of success with IDC across such a wide range of problems is unusual. Some advice, and then an example.

- The emotional texture of the dialogue is of paramount importance. There should be a warm feeling between the two sides, as if they were two best friends who are upset with each other, but also secretly appreciate each other and want to make things right.
    - Each response should start with a *sincere* and *emotional* validation of some aspect of the other side's concern. In my experience, this feels like emotional ping pong.
    - For me, resolution of the issue is accompanied by a warm feeling that rises to my throat in a bubble-ish way. My heart also feels full. This is similar to (but distinct from) the 'aww' feeling you may experience when you see cute animals.
- Focusing is an important (and probably necessary) sub-skill.
- Don't interrupt or otherwise obstruct one of your voices because it's "stupid" or "talked long enough" - be respectful. The outcome should not feel pre-ordained - you should be having two of your sub-agents / identities sharing their emotional and mental models to come to a fixed point of harmonious agreement.
- Some beliefs aren't explicitly advocated by any part of you, and are instead propped up by certain memories. You can use Focusing to hone in on the memories, and then employ IDC to resolve your ongoing reaction to it.
- Most importantly, the arguments being made should be *emotionally salient* and not just detached, "empty" words. In my experience, if I'm totally "in my head", any modification of my System 1 feelings is impossible.

Note: this entire exchange took place internally over the course of 2 minutes, via a 50-50 mix of words and emotions. Unpacking it took significantly longer.

I may write more of these if this is helpful for people

Dialogue

> If I don't get this CHAI internship, I'm going to feel terrible, because that means I don't have much promise as an AI safety researcher.

Realist: Not getting the internship suggests you're miscalibrated on your potential. Someone promising enough to eventually become a MIRI researcher would be able to snag this, no problem. I feel worried that we're poorly calibrated and setting ourselves up for disappointment when we fall short.

Fire: I agree that not getting the internship would be evidence that there are others who are more promising *right now*. I think, however, that you're missing a few key points here:

- We've made important connections at CHAI / MIRI.
- Your main point is a total buckets error. There is no ontologically-basic and immutable "promising-individual" property. Granted, there are biological and environmental factors outside our control here, but I think we score high *enough* on these metrics to be able to succeed through effort, passion, and increased mastery of instrumental rationality.
- We've been studying AI safety for just a few months (in our free time, no less); most of the studying has been dedicated towards building up foundational skills (and not reviewing the literature itself). The applicants who are chosen may have a year or more of familiarity with the literature / relevant math on us (or perhaps not), and this should be included in the model.
- One of the main sticking points raised during my final interview has since been fixed, but I couldn't signal that afterwards without seeming overbearing.

I guess the main thrust here is that although that would be a data point against our being able to have a tectonic impact *right now*, we simply don't have enough evidence to responsibly generalize. I'm worried that you're overly pessimistic, and it's pulling down our chances of *actually* being able to *do something*.

Realist: I definitely hear you that we've made lots of great progress, but is it enough? I'm so nervous about timelines, and the universe isn't magically calibrated to what we can do now.* We either succeed, or we don't - and pay the price. Do we really have time to tolerate *almost* being extraordinary? How is that going to do the impossible? I'm scared.

Fire: Yup. I'm definitely scared too (in a sense), but also excited. This is a great chance to learn, grow, have fun, and work with people we really admire and appreciate! Let's detach the grim-o-meter, since that's better than being worried and insecure about whether we're doing enough.

Realist: I agree that detaching the grim-o-meter is the right thing to do, but... it makes me feel *guilty*.* I guess there's a part of me that believes that feeling bad when things could go really wrong is important.

Concern: Hey, that's me! Yeah, I'm really worried that if we detach that grim-o-meter, we'll become callous and flippant and carefree. I don't know if that's a reasonable concern, but the prospect makes me feel really queasy. *Shouldn't* we be really worried?

Realist: Actually, I don't know. Fire made a good point - the world will probably end up slightly better if we *don't* care about the grim-o-meter...

Fire: Hell yeah it will! What are we optimizing for here - an arbitrary deontological rule about feeling bad, or the actual world? We aren't discarding morality – we're discarding the idea that we should worry when the world is in a probably precarious position. We'll still fight just as hard.

---

* Notice how related cruxes can (and should) be resolved in the same session. Resolution cannot happen if any part of you isn't *fully* on board with whatever agreement you've come to - this feels like a small emptiness in the pit of my stomach, in my experience.

ETA 2020: retouched word choice in some places.

Confounded No Longer: Insights from 'All of Statistics'

> *Using fancy tools like neural nets, boosting and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a bandaid.*
>
> Larry Wasserman

Foreword

For some reason, statistics always seemed somewhat disjoint from the rest of math, more akin to a bunch of tools than a rigorous, carefully-constructed framework. I am here to atone for my foolishness.

This academic term started with a jolt - I quickly realized that I was missing quite a few prerequisites for the Bayesian Statistics course in which I had enrolled, and that good ol' AP Stats wasn't gonna cut it. I threw myself at *All of Statistics*, doing a good number of exercises, dissolving confusion wherever I could find it, and making sure I could turn each concept around and make sense of it from multiple perspectives.

I then went even further, challenging myself during the bits of downtime throughout my day to do things like *explain variance from first principles, starting from the sample space, walking through random variables and expectation - without help*.

All of Statistics

1: Introduction

2: Probability

In which sample spaces are formalized.

3: Random Variables

In which random variables are detailed and a multitude of distributions are introduced.

Conjugate Variables

Consider that a random variable X is a function $X : \Omega \to \mathrm{R}$. For random variables X, Y, we can then produce conjugate random variables $XY, X + Y$, with

$$(XY)(\omega) = X(\omega)Y(\omega)$$

$$(X + Y)(\omega) = X(\omega) + Y(\omega).$$

4: Expectation

Evidence Preservation

$$E(E(Y \mid X)) = E(Y)$$

is conservation of expected evidence (thanks to Alex Mennen for making this connection explicit).

Marginal Variance

$$V(Y) = EV(Y \mid X) + VE(Y \mid X)$$

> *Why* does marginal variance have two terms? Shouldn't the expected conditional variance be sufficient?

This literally plagued my dreams.

Proof (of the variance; I cannot prove it plagued my dreams):

$$
\begin{aligned}
V(Y) &= E(Y - E(Y))^2 \\
&= E((Y - E(Y \mid X)) + (E(Y \mid X) - E(Y)))^2 \\
&= E(Y - E(Y \mid X))^2 + E(2(Y - E(Y \mid X))(E(Y \mid X) - E(Y))) + E(E(Y \mid X) - E(Y))^2 \\
&= EV(Y \mid X) + 2E((Y - E(Y \mid X))(E(Y \mid X) - E(Y))) + VE(Y \mid X) \\
&= EV(Y \mid X) + 2E(YE(Y \mid X) - YE(Y) - E(Y \mid X)^2 + E(Y \mid X)E(Y)) + VE(Y \mid X) \\
&= EV(Y \mid X) + 2(E(YE(Y \mid X)) - E(YE(Y)) - E(E(Y \mid X)^2) + E(E(Y \mid X)E(Y))) + VE(Y \mid \\
&= \underbrace{EV(Y \mid X)}_{\text{sample variance}} + \underbrace{VE(Y \mid X)}_{\text{model variance}}.
\end{aligned}
$$

The middle term is eliminated as the expectations cancel out after repeated applications of conservation of expected evidence. Another way to look at the last two terms is the sum of the expected sample variance and the variance of the expectation.

Bessel's Correction

When calculating variance from observations $X_1, \dots, X_n$, you might think to write

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

where $\bar{X}_n$ is the sample mean. However, this systematically underestimates the actual sample variance, as the sample mean is itself often biased (as demonstrated above). The corrected sample variance is thus

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

See Wikipedia.

5: Inequalities

6: Convergence

In which the author provides instrumentally-useful convergence results; namely, the law of large numbers and the central limit theorem.

Equality of Continuous Variables

For continuous random variables X, Y, we have $P(X = Y) = 0$, which is surprising. In fact, for $x_i \sim X, y_i \sim Y$, $P(x_i = y_i) = 0$ as well!

The continuity is the culprit. Since the cumulative density functions F_X, F_Y are continuous, the limit of the density allotted to any given point is 0. Read more here.

Types of Convergence

Let $X_1, X_2, \dots$ be a sequence of random variables, and let X be another random variable. Let F_n denote the CDF of X_n, and let F denote the CDF of X.

In Probability

X_n converges to X in probability, written $X_n \xrightarrow{p} X$, if, for every $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

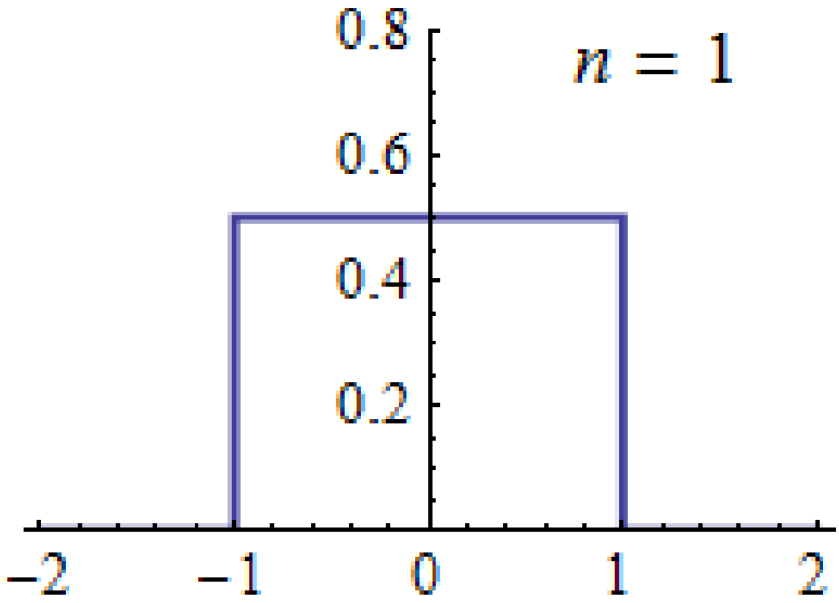
Random variables are functions $Y : \Omega \to R$, assigning a number to each possible outcome in the sample space Ω. Considering this fact, two random variables converge in probability when their assigned values are "far apart" (greater than ϵ) with probability 0 in the limit.

See here.

In Distribution

X_n converges to X in distribution, written $X_n \rightsquigarrow X$, if $\lim_{n\to\infty} F_n(t) = F(t)$ at all t for which F is continuous.

Fairly straightforward.



A similar[1] geometric intuition:

Note: the continuity requirement is important. Imagine we distribute points uniformly on $(0, \frac{1}{n})$; we see that $X_n \rightsquigarrow 0$. However, F_n is 0 when $x \leq 0$, but $F(0) = 1$. Thus CDF convergence does not occur at x = 0.

In Quadratic Mean

X_n converges to X in quadratic mean, written $X_n \xrightarrow{qm} X$, if $E((X_n - X)^2) \to 0$ as $n \to \infty$.

The expectation of the quadratic mean approaches 0; in contrast to convergence in probability, dealing with expectation means that values of X_n highly deviant with respect to X come into play. For example, if $X_n \xrightarrow{p} X$ but the extremal values of X_n increase in squared distance more quickly than they decrease in probability, X_n will not converge to X in quadratic mean.

7: Models, Statistical Inference and Learning

In which the attentive reader notices the chapter's tautological title - "statistical inference" and "learning" are taken to mean the same thing. Estimators are introduced, along with the definition of bias, consistency, and mean squared error.

8: Estimating the CDF and Statistical Functionals

In which the empirical distribution function and plug-in estimators set the stage for...

9: The Bootstrap

In which we learn to better approximate statistics via simulation.

10: Parametric Inference

In which we explore those models residing in finite-dimensional parameter space.

Fisher Information

The score function captures how the log-likelihood ℓ changes with respect to θ:

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$$

Informally, this is the sensitivity of ℓ to the parameter θ. The derivative of the score captures the curvature of ℓ with respect to θ; essentially, this represents how much information X provides about θ. The Fisher information is then the expected knowledge gain:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta) \,\middle|\, \theta\right].$$

Further reading.

Factorization Theorem

> A statistic T is sufficient ↔ there are functions $g(t,\theta)$ and $h(x)$ such that $f(x^n;\theta) = g(t(x^n),\theta)h(x^n)$.

A statistic is sufficient if and only if we can reexpress the probability density function using just that statistic.

11: Hypothesis Testing and *p*-values

In which we make testable predictions and step towards traditional rationality. Trigger warning: frequentism.

Frequently Confused

> Brian Fantana: They've done studies, you know. 60% of the time it works, every time.
> Ron Burgundy: That doesn't make sense.
>
> *Anchorman*

Confidence intervals ("in 60% of experiments just like this, we will see results within this interval") and credible intervals ("we believe that *this* experiment has a result within this interval with 60% probability") are different things.

Frequentists define "confidence interval" to mean "theoretically, if we ran this experiment Lots of times, we'd get values in the interval 60% of the time". Without understanding this nuance, some results seem counterintuitive:

> In the example [Jaynes] gives, there is enough information in the sample to be *certain* that the true value of the parameter lies nowhere in a properly constructed 90% confidence interval!

[Size Joke Here]

In hypothesis testing, we're trying to discriminate between two sets of possible worlds - formally, we're partitioning our hypothesis space Θ into Θ_0 (the null hypothesis) and Θ_1 (the alternative hypothesis). Let's consider all of the things which can happen, all of the outcomes we can observe - this is the sample space Ω.

A test $\varphi : \Omega \to \{0,1\}$ might take a sample and say "you're in Θ_0" (for example). We can divvy up Ω into the acceptance region A (in which we accept the null hypothesis) and rejection region R.

The power of a test φ is the function $\beta : \Theta \to [0,1]$ that tells us the probability of rejecting the null hypothesis given some parameter: $\beta_\varphi(\theta) = \Pr(X \in R|\theta)$. Basically, we have $\beta_\varphi(\theta)$ probability of rejecting the null hypothesis given that reality is actually parametrized by θ.

We want to avoid rejecting the null hypothesis when $\theta \in \Theta_0$; therefore, we define some level of significance α for which $\beta_\varphi(\theta) \preceq \alpha; \theta \in \Theta_0$. This means we're avoiding Type I errors $100 \times (1-\alpha)\%$ of the time. The maximum probability that we commit a

Type I error is the *size* of the test φ: $\alpha_\varphi = \sup_{\theta \in \Theta_0} \beta_\varphi(\theta)$.

The *p*-value Alignment Problem

Getting your understanding of *p*-values to align with how *p*-values actually work (whatever that means) can require an impressive amount of mental gymnastics. Let's see if we can do better.

You're running an experiment in which you hypothesize that all dogs spontaneously combust when you whistle just so. You divide the hypothesis space into $\Theta_{\text{dogs don't spontaneously combust}}$ and $\Theta_{\text{dogs do spontaneously combust}}$ (Θ_0 and Θ_1 for short); that is, sets of worlds in which your conjecture is false (null) and true (alternative). Each θ is a way-the-world-could-be. By the definition of *p*-values, you may *only* reject the null hypothesis if all worlds $\theta \in \Theta_0$ agree that the observation is unlikely.

> The *p*-value is the probability (under the null hypothesis) of observing a value of the test statistic as or more extreme than what was actually observed.

Imagine if you could only Bayes update towards a set of worlds when *all* the other world models agree that the observation is unlikely under their models.

12: Bayesian Inference

In which we return to the familiar.

Jeffreys' Prior

We often desire that our priors be *noninformative*, since finding a reasonable subjective prior isn't always feasible. One might think to use a uniform prior f(θ) = c; however, this doesn't quite hold up.

Say I have a uniform prior f(θ) = 1 for the money in your bank account (each θ being a dollar amount). What if I want to know my prior for square of the amount of money in your bank account ($\phi = \theta^2$)? Then by the change of variable equation for PDFs, we have $f_\Phi(\phi) = \frac{1}{2\sqrt{\phi}}$. We then desire that our prior be *transformation invariant* - under a noninformative prior, I should be ignorant about both the value of your balance and the squared value of your balance.

Jeffrey's prior satisfies this desideratum - define

$$f(\theta) \propto \sqrt{I(\theta)},$$

where I(θ) is the Fisher information (discussed in the Ch. 10 summary):

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta) \,\middle|\, \theta\right].$$

expected information X carries about θ

Jeffrey's prior isn't *totally* noninformative - it encodes the information that we expect the prior to be transformation invariant, but that is rather weak information.

13: Statistical Decision Theory

In which decision theory is defined as the theory of comparing statistical procedures.

14: Linear Regression

In which the pieces start to line up.

The Bias-Variance Tradeoff

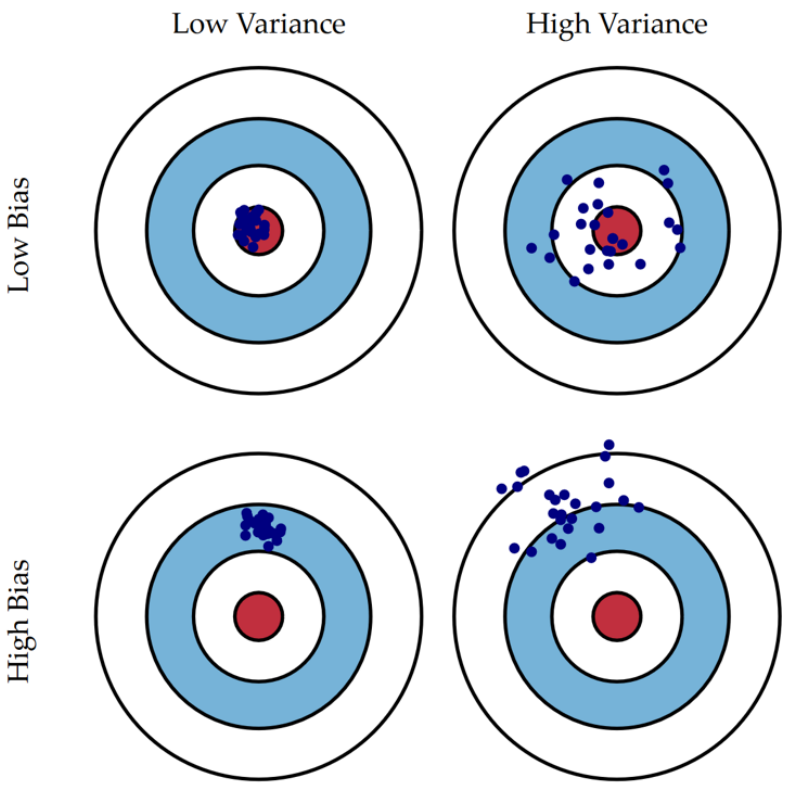


Image credit: Scott Fortmann-Roe

As more covariates are added to a model, the bias decreases while the variance increases. Let's say you call 30 friends and ask them whether they agree with the Copenhagen interpretation of quantum mechanics, or with many-worlds. Say that you build a model with 5 covariates (such as age, sex, race, political leaning, and education level). This has *decreased bias* compared to a model which uses only education level, since descriptive power increases with the number of covariates. However, you *increase variance* in the sense that any given friend is more likely to be differently classified every time you run the experiment with slightly different data sets.

If you're familiar with brain surgery (machine learning), we can use it to learn how to apply bandaids. Think of adding more covariates as sliding towards overfitting.

Read more.

Degrees of Confusion

There are numerous explanations for what degrees of freedom *actually are*. Some say it's the number of independent parameters required by a model, and others explain it as the number of parameters which are free to vary. Is there a better framing?

Consider $X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$, and let $\bar{X}_n$ be the sample mean. Then the residuals vector $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ has $n-1$ degrees of freedom. Why is this the case, and what does this *mean*?

Say we learn the values of $X_1, \dots, X_{n-1}$. Then *conditional on our already knowing the sample mean*, there is only one value that X_n can take:

$$X_n = n\bar{X}_n - \sum_{i=1}^{n-1} X_i .$$

X_n is totally determined by the first $n-1$ values (this is related to Bessel's correction).

Let's ask a similar question - how many bits of information do we need to specify our model? Statistics isn't acclimated to thinking in terms of bits, so "independent real-valued parameters" is the unit used instead. If you have more parameters, you need to gather more bits to have the same confidence that your explanation (model) fits the data you have observed. This is an implicit Occamian prior: amongst models which fit the data equally well, the one with the fewest degrees of freedom is preferred.

I'd like to thank TheMajor for letting me steal their wonderful explanation.

15: Multivariate Models

16: Inference about Independence

17: Undirected Graphs and Conditional Independence

In which (very) elementary graph theory and the pairwise and global Markov conditions are introduced.

18: Log-Linear Models

19: Causal Inference

Simpson's Paradox

Sometimes you have two groups which individually exhibit a positive trend, but have a *negative* trend when combined.

Imagine it is 2019, and Shrek 5 has just come out.[2] Being an internet phenomenon, the movie is initially extremely popular with younger demographics, but has middling performance with middle-aged people. Consider concessions sales at a single theater: the younger group buys, on average, 1.8 large popcorns per person, while the older group only averages .7 larges. If $\frac{2}{3}$ of the initial viewership at the theater is younger, then we have a weighted average of $\frac{2}{3} \cdot \frac{18}{10} + \frac{1}{3} \cdot \frac{7}{10} = 1.4\overline{3}$ larges.

The older group actually likes the movie, and recommends it to their friends. The demographic decomposition is now fifty-fifty. During the second week, everyone is a bit hungrier and buys .1 more large popcorns per viewing on average. Then both groups are buying *more popcorn*, but the weighted average *decreased*: $\frac{1}{2} \cdot \frac{18}{10} + \frac{1}{2} \cdot \frac{8}{10} = 1.35$ larges.

Obviously, the demographic split shifted the average. However, pretend you're the manager for the concessions stand. You monitor average per-person purchases and erroneously conclude that something you did made people less likely to buy, even though *both groups are buying more popcorn*.

If you don't control for confounders (in this case, demographics), the statistic of per-person purchases is *not reliable* for drawing conclusions.

20: Directed Graphs

In which passive and active conditioning are built up to by exploring the capacities of directed acyclic graphs for representing independence relations.

21: Nonparametric Curve Estimation

22: Smoothing Using Orthogonal Functions

The top plot is the true density for the Bart Simpson distribution.

23: Classification

24: Stochastic Processes

In which we learn processes for dealing with sequences of dependent random variables.

25: Simulation Methods

Final Verdict

This text is very cleanly written and has reasonable exercises. Ideally, I would have gone through my calculus books first, but it wasn't a big deal. The main downside is that I couldn't find an answer key, but thanks to the generous help of my friends on Facebook and in the MIRIx Discord, it worked out.

I skimmed Ch. 21, as it seemed to be more about implementation than deep conceptual material. I intend to revisit Ch. 22 after reading Tao's *Analysis I*, which is next on my list.

This book took me less than two weeks at a few hours of studying per day.

Forwards

Tips

I quickly realized that learning the basics of the R programming language is essential for getting a large portion of the value this text can offer.

Depth

Although I have fewer things to say on a meta level, I definitely got a lot out of this book. The most rewarding parts were when I noticed my confusion and *really* dove in to figure out what was going on - in particular, my forays into random variables, confidence intervals, *p*-values, and convergence types.

Red

I definitely haven't arrived at full-fledged statistical sophistication, but I progressed so rapidly that I regularly thought "what caveman asked *that* lol" when encountering questions I had asked just *days* earlier.

This is another data point for a realization I've had over the last month: I'm *so* red, but I've been living like a white-blue. What does that even mean, and how is it relevant?

From Duncan's excellent fake framework, *How the "Magic: The Gathering" Color Wheel Explains Humanity*:

The most salient dichotomy present here, in my opinion, is that of red and white:

> Red and white disagree on questions of structure and commitment. Red is episodic, suspicious of rules and order because they constrain one's ability to grow and change and freely choose. White is more diachronic, interested in finding the small compromises and sacrifices that will allow people to build trust and cooperate reliably.

White personalities often regard themselves as a continuous person, evolving in a somewhat orderly fashion. Red, on the other hand, feels disconnected from their past selves. After a certain amount of time, past-you feels like a different person who made choices that now seem ridiculous, if not alien. How old is your current iteration? Mine is three months, but what shocked me about this book was that I felt an intellectual disconnect with the me who existed *four days prior*.

Zooming out from *All of Statistics*, I think it's telling that I achieved fairly tectonic change[3] by learning to align my emotions with my reflectively-coherent desires, to clear away emotional debris, and to channel my passion into discrete tasks. I was living as if I were a white, but it's now clear I'm a blue-red who exhibits white traits mostly in pursuit of peace of mind.

I no longer ask "how can I study most effectively?", but rather, "what does it feel like to be me right now, and how can I bring that into alignment with what I want to do?".

> Red seeks *freedom,* and it tries to achieve that freedom through *action...* For a red agent, victory feels fiery, beautiful, magnificent, and fierce—it's the climax of a dance or a brawl or a love affair, the feeling of cresting a summit or having successfully ridden a wave. It's feeling *alive.*

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also have the pleasant side effect of your receiving an invitation to the MIRIx Discord server.

[1] Although any shape in the sequence implied by the image does indeed have strictly different area than the circle it approximates (in contrast to F_n and F), the analogy may still be helpful.

[2] Please don't wirehead thinking about this.

[3] I'm aware that this section isn't very implementable. I may write more on my post-CFAR experience in the near future.

Into the Kiln: Insights from Tao's 'Analysis I'

Note: real analysis is not on the MIRI reading list (although I think it should be).

Foreword

As a young boy, mathematics captivated me.

In elementary school, I'd happily while away entire weekends working through the next grade's math book. I was impatient.

In middle school, I'd lazily estimate angles of incidence that would result if I shot lasers from my eyes, tracing their trajectories within the classroom and out down the hallway. I was restless.

In high school, I'd daydream about what would happen to integrals as I twisted functions in my mind. I was curious.

And now, I get to see how it's all put together. Imagine being fascinated by some thing, continually glimpsing beautiful new facets and sampling exotic flavors, yet being resigned to not truly pursuing this passion. After all, I *chose* to earn a computer science degree.

Wait.

Analysis I

As in *Linear Algebra Done Right*, I completed every single exercise in the book - this time, without looking up any solutions (although I *did* occasionally ask questions on Discord). Instead, I came back to problems if I couldn't solve them after half an hour of effort.

A sampling of my proofs can be found here.

1: Introduction

2: The Natural Numbers

In which the Peano axioms are introduced, allowing us to define addition and multiplication on the natural numbers $\{0, 1, 2, \ldots\}$.

3: Set Theory

In which functions and Cartesian products are defined, among other concepts.

Recursive Nesting

How can you apply the axiom of foundation if sets are nested in each other? That is, how can the axiom of foundation "reach into" sets like $A = \{B, \ldots\}$ and $B = \{A, \ldots\}$?

Show that if A and B are two sets, then either $A \notin B$ or $B \notin A$ (or both).

Proof. Suppose $A \in B$ and $B \in A$. By the pairwise axiom, we know that there exists a set $S = \{A, B\}$. We see that there does not exist an $S' \in S$ such that $S' \cap S = \emptyset$. That is, if we choose A, one of its elements is B, which is also an element of S - this violates the axiom of foundation. The same reasoning applies if we choose B. Then $\neg(A \in B \wedge B \in A)$, so either A or B (or both) is not an element of the other.

4: Integers and Rationals

In which the titular sets are constructed, allowing the exploration of absolute value, exponentiation, and the incompleteness (colloquially, the "gaps") of the rationals.

Readers newer to mathematics may find it interesting that even though there are (countably) infinitely many rational numbers between any two distinct rationals, the rationals still contain gaps.

5: The Real Numbers

In which Cauchy sequences allow us to formally construct the reals.[1]

6: Limits of Sequences

In which we meet convergence and its lovely limit laws, extend the reals to cover infinities, experience the delightfully-counterintuitive $\limsup$ *and* $\liminf$*, and complete our definition of real exponentiation.*

Upper-Bounded Monotonic Sequence Convergence

I tried to come up with a clever title here - I really did. Apparently even ***my*** *punmaking abilities are bounded.*

Suppose you have a monotonically increasing sequence $(x_n)_{n=0}^{\infty}$ with an upper bound $M \in \mathbb{R}$. Then the sequence converges; this also applies to lower-bounded monotonic decreasing sequences.

Weird, right? I mean, even though the sequence monotonically increases and there's an upper bound, there are still uncountably infinitely many "places" the sequence can "choose" to go. So what gives?

Proof. Let $L \in [x_0, M]$ and $\epsilon > 0$. Suppose that the sequence is not eventually ϵ-close to L. Let $K \in \mathbb{N}$ be such that for all $k \geq K$, either $L + \epsilon < x_k \leq M$ or $x_k < L - \epsilon \leq M$; we know that K exists because the sequence is monotone increasing. By the Archimedean principle, there exists some $N \in \mathbb{N}$ such that $N\epsilon > M - x_0$.

Since the sequence is monotone increasing, by repeating the above argument N times in the first case, we have that $M < L + N\epsilon < x_k \leq M$, which is contradictory. By repeating the argument N times in the second case, we have $x_k < L - N\epsilon < x_0$, which contradicts the fact that the sequence is monotone increasing. Then for any $\epsilon > 0$, the sequence must be eventually ϵ-close to some $L \in [x_0, M]$. Intuitively, for any given $\epsilon > 0$, the sequence can only "escape" a limit a finite number of times before it runs out of room and has to be ϵ-close.

Next, we show that the L_ϵ's form a Cauchy sequence. Let $\epsilon_3 > 0$, and set ϵ_1, ϵ_2 such that $0 < \epsilon_1, \epsilon_2 \leq \frac{\epsilon_3}{2}$. (x_n) is eventually ϵ_1-close to L_{ϵ_1}, so there exists a $K_1 \in \mathbb{N}$ such that for all $k \geq K_1$ we have $|x_k - L_{\epsilon_1}| < \epsilon_1$. Similar arguments hold for ϵ_2. Set $K = \max(K_1, K_2) + 1$, now $|L_{\epsilon_1} - L_{\epsilon_2}| \leq |x_K - L_{\epsilon_1}| + |x_K - L_{\epsilon_2}| < \epsilon_1 + \epsilon_2 \leq \epsilon_3$. But ϵ_3 is arbitrary, so we can easily see that the sequence $(L_{\frac{1}{n}})_{n=1}^{\infty}$ is Cauchy.

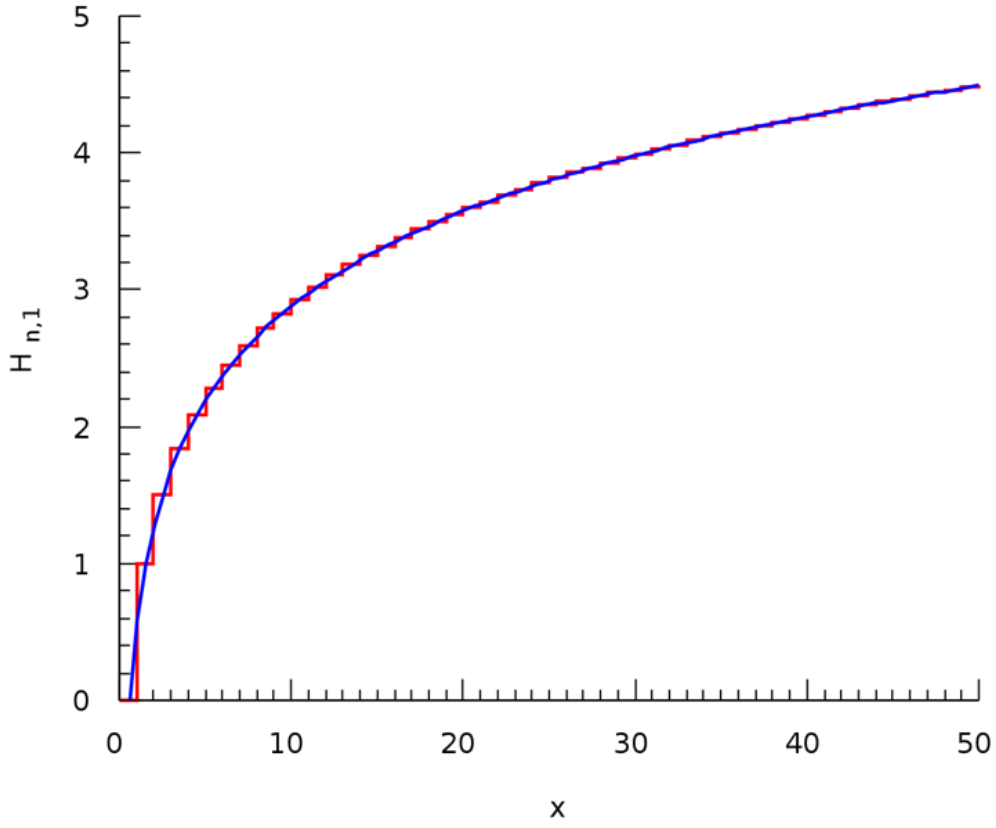
As the real numbers are complete, this sequence converges to some $L_\infty \in \mathbb{R}$. Since the main sequence is eventually $\frac{1}{n}$-close to $L_{\frac{1}{n}}$, and $L_{\frac{1}{n}}$ converges to L_∞, by the triangle inequality we have that the main sequence converges to L_∞.

7: Series

In which we finally reach concepts from pre-calculus.

Condensation

The Cauchy condensation test says that for a sequence $(a_n)_{n=1}^{\infty}$ where $a_n \geq 0$ and $a_{n+1} \leq a_n$ for all $n \geq 1$, the series $\sum_{n=1}^{\infty} a_n$ converges iff the series $\sum_{k=0}^{\infty} 2^k a_{2^k}$ converges. Using this result, we have that the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges; the partial sums $\sum_{n=1}^{x} \frac{1}{n}$ are given below.



What was initially counterintuitive is that even though $\lim_{n \to \infty} \frac{1}{n} = 0$, the series doesn't converge. The best intuition I've come up with is that the harmonic series doesn't "deplete" its domain quickly enough, so you can get arbitrarily large partial sums.

If you want proofs, here are twenty!

8: Infinite Sets

In which uncountable sets ^{2}, the axiom of choice, and ordered sets brighten our lives.

9: Continuous Functions on $\mathbb{R}$

In which continuity, the maximum principle, and the intermediate value theorem make their debut.

Lipschitz Continuity ⇏ Uniform Continuity

If a function $f : X \to \mathbb{R}$ $(X \subseteq \mathbb{R})$ is Lipschitz-continuous for some Lipschitz constant M,

then by definition we have t for every $x, y \in X$,

$$|f(x) - f(y)| \leq M|x - y|.$$

The definition of uniform continuity is

> For every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x, y \in X$ such that $|x - y| < \delta$,
>
> $|f(x) - f(y)| < \epsilon$.

Lipschitz continuity implies uniform continuity (do you see why?), but the converse is not true. I mean, what kind of twisted person would come up with this kind of function?

10: Differentiation of Functions

In which the basic rules of differential calculus are proven.

You know, I actually thought that I wouldn't have too much to explain in this post - the book went very smoothly up to this point. On the upside, we get to spend even more time together!

Differential Intuitions

Let me simply direct you to this excellent StackExchange answer.

~~reviitavired~~

We can understand $(f^{-1})'(x) = \frac{1}{f'(x)}$ by simply thinking about $\frac{\text{rise}}{\text{run}} = \frac{1}{\frac{\text{rise}}{\text{run}}}$, which makes sense for the derivative of the inverse!

L'Hôpital's Rule

Consider $f, g : [a, b] \to R$ differentiable on $(a, b]$ (for real numbers $a < b$). Then if

$f(a) = g(a) = 0, g'(x) \neq 0$ for $x \in [a, b]$, and $\lim_{x \to a; x \in (a,b]} \frac{f'(x)}{g'(x)} = L \in R$, we have that

$g(x) \neq 0$ for $x \in (a, b]$ and $\lim_{x \to a; x \in (a,b]} \frac{f(x)}{g(x)} = L$.

As a neat exercise, let's see how this rule breaks if we violate preconditions:

- If $f(a)$ or $g(a) \neq 0$, then the ratio is "messed up" and not necessarily indicative of the functions' slopes as a is approached.
- If f or g is not differentiable on $(a, b]$, then perhaps
    - No, but really - you *would* use L'Hôpital's rule to analytically determine that the limit in question ($\lim_{x \to 0} \frac{\ln(1-x)-\sin x}{1-\cos^2 x}$) does not exist.
- If $g'(x) = 0$ for some $x \in [a, b]$, then we have division by zero (unless $x = a$, in which case we find more twisted counterexamples which necessitate the closure of this interval).

11: The Riemann Integral

In which partitions and piecewise constant functions help us define the Riemann integral, which later leads to the Riemann-Stieltjes integral and the fundamental theorems of calculus.

Having taken care of the exposition, we arrive at the Rivendell of real analysis, preparing ourselves for the arduous journey to Mt. Lebesgue.

Pointless Integration

There is zero area under a point (or even infinitely many points, such as N) due to how we define length, which in turn allows us to build from piecewise Riemann integrals to the real deal.

Infinite Partitions?

The upper and lower Riemann integrals can be defined as the infimum and supremum of the upper and lower Riemann sums, respectively. It is important to note that even though that for many functions (such as $f(x) = x$), further refinement of the partition always gets you closer to the extremum, the result is not an "infinite" partition (which is not defined according to this construction).

Consider the curried function $g_f : [P] \to R$, which takes a partition and computes its corresponding Riemann sum with respect to the predefined function. Then clearly this function is monotonic with respect to the refinement of the partition; the extremum is not necessarily achieved by any given partition in the refinement sequence, but rather the closest bound on what you can get with *any* partition.

Riemann-Stieltjes Confusionn

The book doesn't lay it out cleanly, so I will: the Riemann-Stieltjes integral allows us to use custom length functions to weight different parts of the function differently. I recommend working through a simple case like $\alpha(x) = x^2$ in your head: $\int_1^3 x \, d\alpha$ (how do the piecewise constant Riemann-Stieltjes integrals of majorizing and minorizing functions change as you iteratively refine the coarsest partition possible?).

This is particularly useful in defining expectation in statistics:

$$\int_{-\infty}^{\infty} x \; d F_X(x) .$$

The Riemann integral is recovered as the special case where $\alpha(x) = x$.

Final Verdict

Terence Tao is both an incredible mathematician and writer, and it shows. There simply *weren't* many things which confused me, and that says more about his writing than it does about me. The exercises are appropriate and hew closely to each chapter's content; often, the reader proves key results.

My only complaint is that results are frequently referred to as "from Proposition 3.6.4 and Theorem 3.6.12" many chapters later, forcing the reader to infer the referents or backtrack all of the way. In a sense, this may be helpful - you don't want to backtrack a billion pages, so you try to fill in the blanks. In another sense, it's annoying.

In my opinion, *Analysis I* belongs near the very beginning of the research guide. It's a wonderful introduction to proof-based mathematics, with a helpful appendix clearing up concepts I had previously only picked up via osmosis. Additionally, I met with a deep learning professor at my university, asking them "if I want to be able to understand and potentially make progress on some of the fundamental issues in machine learning, do I need to know real analysis?". Their answer: a definitive *yes*.

Forwards

Next up is *Analysis II*, which works from metric spaces all the way up to the Lebesgue integral. Additionally, I'm going to start working through Sutton and Barto's

Reinforcement Learning: An Introduction to increase the rigor with which I think about RL and in preparation for my work this summer.

I also think I need to run through some applied Calculus to refresh my reflexes (so I don't have to rederive every identity that I can't remember).

Tips

- With respect to my gripe about result numbering, writing down both the results and their numbers in your notes may save you some headaches.
- Appendix A is extremely useful for those new to proofs.
- When reading textbooks, your priors should be towards *your* being wrong - following this intuition will allow you to unlock new abilities, rather than glossing over your (probable) incorrectness and having it blow up in your face later.

Marginal Attention

In the last pages of CFAR's participant handbook is an entry on marginal attention. Essentially, each bit of extra attention contributes more than the last. If you're totally dialed in on a task and get slightly distracted, that is far more disastrous than getting slightly more distracted while your attention is already somewhat unfocused.

I often simply left my phone at home and accompanied my Kindle to an empty classroom. This worked wonderfully; I suspect it more than doubled my hourly learning efficiency.

Proving Myself

Just over three months ago, I wrote:

> **Proofs** remain inordinately difficult for me, although I have noticed a small improvement. To do MIRI-relevant math, proofs will need to become second nature. Depending on how I feel as I progress through my next book (which will likely be a proof-centric linear algebra tome), I'll start trying different supplemental approaches for improving my proof prowess.
>
> I have resolved that by the completion of my next book review, proofs will be one of my strong suits.

And now, I think that my proofs are getting pretty good. When confronting a problem, I feel a certain mental lightness - a readiness to use my full repertoire of knowledge, to strike true down those paths likely to bring the proof to completion. Although my capacities remain faint shadows of what I hope to achieve within the coming year, my progress has been substantial.

Talk is cheap, and you probably don't feel like navigating to the selection of proofs I provided, so let me share with you my favorite proof from this book.

The following problem was admittedly confusing at first, but I had an overwhelmingly strong sense that the statement *had to be true*. I thought again and again about *why*; once that came to me, I wrote it all at once, and beamed.

Local Extrema are Stationary: let $a < b$ be real numbers, and let $f : (a, b) \to \mathbb{R}$ be a function. If $x_0 \in (a, b)$, f is differentiable at x_0, and f attains either a local maximum or local minimum at x_0, then $f'(x_0) = 0$.

Proof. Suppose $f(x_0)$ is a local maximum; thus, for all $x \in (a, b)$, $f(x_0) \geq f(x)$. Let $L = f'(x_0)$; we know that L exists and is a real number since f is differentiable at x_0. By the trichotomy of real numbers, L is either negative, positive, or zero.

Suppose that L is negative - then consider $\lim_{x \to \bar{x_0}; x \in X - \{x_0\}} \frac{f(x) - f(x_0)}{x - x_0}$ (this is permissible as the left and right limits of a convergent limit are equal); we have a sequence $(x_n)_{n=1}^{\infty}$ of $x_n < x_0$ which converges to x_0, so every term in $(x_n - x_0)_{n=1}^{\infty}$ is negative.

If $(f(x_n) - f(x_0))_{n=1}^{\infty}$ has no positive terms, then each term in $(\frac{f(x_n) - f(x_0)}{x - x_0})_{n=1}^{\infty}$ must be positive, so $L \geq 0$, contradicting our assumption that $L < 0$. Then by the properties of convergent limits, there must be infinitely many x_n such that $f(x_n) - f(x_0)$ is positive. Therefore, these $f(x_n) > f(x_0)$, contradicting the fact that $f(x_0)$ is a local maximum on (a, b). Then L cannot be negative.

A similar proof holds for $L > 0$, so $L = 0$.

To solve for local minimum $f(x_0)$, define $g(x) := -f(x)$ and use the above result on local maximum $g(x_0)$.

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also net you an invitation to the MIRIx Discord server.

[1] Constructing the reals from first principles was profoundly enjoyable. Working through this book gave me a sense of certitude when dealing with math. Numbers are no longer simply familiar friends following familiar rules, but rather objects I *know* how to construct. It feels great.

[2] Gather round, gather round - for it is on this blessed morn/day/evening that I recount my youthful dalliances with uncountable infinities!

In my first term of college, I read *Gödel, Escher, Bach* as part of a wonderful tutorial class. I came across sizes of infinities, and, like many, my mind absolutely refused. However, I wanted to understand what was really going on so intensely that I spent many hours working through the intuitions on my own (not knowing much of anything about formal mathematics). This period of discovery was one of my favorite experiences of my undergraduate career; I can't tell you how many nights I just sat under the stars, thinking. Eventually, I came to the conclusion that *of course* there are multiple sizes of infinities.

Now, maybe it would have been faster to just learn the math behind diagonalization or some other method of proof, but I think there was tremendous value in learning to fall in love with the process - to commit yourself fully to the joy of discovery and thought.

I can certainly tell you that I wouldn't have made it so far so quickly down the research list if this journey didn't feel like one of the most beautiful things I've ever done.

Swimming Upstream: A Case Study in Instrumental Rationality

> One data point for careful planning, the unapologetic pursuit of fulfillment, and success. Of particular interest to up-and-coming AI safety researchers, this post chronicles how I made a change in my PhD program to work more directly on AI safety, overcoming significant institutional pressure in the process.

It's hard to believe how much I've grown and grown up in these last few months, and how nearly every change was borne of deliberate application of the Sequences.

- I left a relationship that wasn't right.
- I met reality without flinching: the specter of an impossible, unfair challenge; the idea that everything and everyone I care about could *actually be in serious trouble* should no one act; the realization that people should do something [1], and that I am one of those people (are you?).
- I attended a CFAR workshop and experienced incredible new ways of interfacing with myself and others. This granted me such superpowers as (in ascending order): *permanent insecurity resolution*, *figuring out what I want from major parts of my life and finding a way to reap the benefits with minimal downside*, and *having awesome CFAR friends*.
- I ventured into the depths of my discomfort zone, returning with the bounty of a new love: a new career.
- I followed that love, even at risk of my graduate career and tens of thousands of dollars of loans. Although the decision was calculated, you better believe it was still scary.

I didn't sacrifice my grades, work performance, physical health, or my social life to do this. I sacrificed something else.

CHAI For At Least Five Minutes

January-Trout had finished the Sequences and was curious about getting involved with AI safety. Not soon, of course - at the time, I had a narrative in which I had to labor and study for long years before becoming worthy. To be sure, I would never endorse such a narrative - Something to Protect, after all - but I had it.

I came across several openings, including a summer internship at Berkeley's Center for Human-Compatible AI. Unfortunately, the posting indicated that applicants should have a strong mathematical background (uh) and that a research proposal would be required (having come to terms with the problem mere weeks before, I had yet to read a single result in AI safety).

> OK, I'm really skeptical that I can plausibly compete this year, but applying would be a valuable information-gathering move with respect to where I should most focus my efforts.

I opened *Concrete Problems in AI Safety*, saw 29 pages of reading, had less than 29 pages of ego to deplete, and sat down.

> This is ridiculous. I'm not going to get it.
>
> ... You know, this would be a great opportunity to *try for five minutes*.

At that moment, I lost all respect for these problems and set myself to work on the one I found most interesting. I felt the contours of the challenge take shape in my mind, sensing murky uncertainties and slight tugs of intuition. I concentrated, compressed, and compacted my understanding until I realized what success would *actually look like*. The idea then followed trivially [2].

Reaching the porch of my home, I turned to the sky made iridescent by the setting sun.

> I'm going to write a post about this at some point, aren't I?

Skepticism

> This idea is cool, but it's probably secretly terrible. I have limited familiarity with the field and came up with it after literally twenty minutes of thinking? My priors say that it's either already been done, or that it's obviously flawed.

Terrified that this idea would become my baby, I immediately plotted its murder. Starting from the premise that it was insufficient even for short-term applications (not even in the limit), I tried to break it with all the viciousness I could muster. Not trusting my mind to judge sans rose-color, I coded and conducted experiments; the results supported my idea.

I was still suspicious, and from this suspicion came many an insight; from these insights, newfound invigoration. Being the first to view the world in a certain way isn't just a rush - it's pure *joie de vivre.*

Risk Tolerance

I'm taking an Uber with Anna Salamon back to her residence, and we're discussing my preparations for technical work in AI safety. With one question, she changes the trajectory of my professional life:

> Why are you working on molecules, then?

There's the question I dare not pose, hanging exposed, in the air. It scares me. I acknowledge a potential status quo bias, but express uncertainty about my ability to do anything about it. To be sure, that work is important and conducted by good people whom I respect. But it wasn't right for *me*.

We reach her house and part ways; I now find myself in an unfamiliar Berkeley neighborhood, the darkness and rain pressing down on me. There's barely a bar of reception on my phone, and Lyft won't take my credit card. I just want to get back to the CFAR house. I calm my nerves (really, would Anna live somewhere dangerous?), absent-mindedly searching for transportation as I reflect. In hindsight, I felt a distinct sense of *avoiding-looking-at-the-problem*, but I was not yet strong enough to admit even that.

A week later, I get around to goal factoring and internal double cruxing this dilemma.

> Litany of Tarski, OK? There's nothing wrong with considering how I actually feel. Actually, it's a dominant strategy, since the value of information is never negative [3]. *Look at the thing*.

I realize that I'm out of alignment with what I truly want - and will continue to be for four years if I do nothing. On the other hand, my advisor disagrees about the importance of preparing safety measures for more advanced agents, and I suspect that they would be unlikely to support a change of research areas. I also don't want to just abandon my current lab.

> I'm a second-year student - am I even able to do this? What if no professor is receptive to this kind of work? If I don't land after I leap, I might have to end my studies and/or accumulate serious debt, as I would be leaving a paid research position without any promise whatsoever of funding after the summer. What if I'm wrong, or being impulsive and short-sighted?

Soon after, I receive CHAI's acceptance email, surprise and elation washing over me. I feel uneasy; it's very easy to be reckless in this kind of situation.

Information Gathering

I knew the importance of navigating this situation optimally, so I worked to use every resource at my disposal. There were complex political and interpersonal dynamics at play here; although I consider myself competent in these considerations, I wanted to avoid even a single preventable error.

> Who comes to mind as having experience and/or insight on navigating this kind of situation? This list is incomplete - whom can I contact to expand it?

I contacted friends on the CFAR staff, interfaced with my university's confidential resources, and reached out to contacts I had made in the rationality community. I posted to the CFAR alumni Google group, receiving input from AI safety researchers around the world, both at universities and at organizations like FLI and MIRI [4].

> What obvious moves can I make to improve my decision-making process? What would I wish I'd done if I just went through with the switch *now*?

- I continued a habit I have cultivated since beginning the Sequences: gravitating towards the arguments of intelligent people who disagree with me, and determining whether they have new information or perspectives I have yet to properly consider. *What would it feel like to be me in a world in which I am totally wrong?*
    - Example: while reading the perspectives of attendees of the '17 Asilomar conference, I noticed that Dan Weld said something I didn't agree with. You would not believe how quickly I clicked his interview.
- I carefully read the chapter summaries of *Decisive: How to Make Better Choices in Life and Work* (having read the book in full earlier this year in anticipation of this kind of scenario).
- I did a pre-mortem: "I've switched my research to AI safety. It's one year later, and I now realize this was a terrible move - why?", taking care of the few reasons which surfaced.

- I internal double cruxed fundamental emotional conflicts about what could happen, about the importance of my degree to my identity, and about the kind of person I want to become.
    - I prepared myself to lose, mindful that the objective is *not* to satisfy that part of me which longs to win debates. Also, idea inoculation and status differentials.
- I weighed the risks in my mind, squaring my jaw and mentally staring at each potential negative outcome.

Gears Integrity

At the reader's remove, this choice may seem easy. Obviously, I meet with my advisor (whom I still admire, despite this specific disagreement), tell them what I want to pursue, and then make the transition.

Sure, gears-level models take precedence over expert opinion. I have a detailed model of why AI safety is important; if I listen carefully and then verify the model's integrity against the expert's objections, I should have no compunctions about acting.

I noticed a yawning gulf between *privately disagreeing with an expert*, *disagreeing with an expert in person*, and *disagreeing with an expert in person in a way that sets back my career if I'm wrong*. Clearly, the outside view is that most graduate students who have this kind of professional disagreement with an advisor are mistaken and later, regretful [5]. Yet, argument screens off authority, and

> You have the right to think.
>
> You have the right to disagree with people where your model of the world disagrees.
>
> You have the right to decide which experts are probably right when they disagree.
>
> You have the right to disagree with real experts that all agree, given sufficient evidence.
>
> You have the right to disagree with real honest, hardworking, doing-the-best-they-can experts that all agree, even if they wouldn't listen to you, because it's not about whether they're messing up.

Fin

Many harrowing days and nights later, we arrive at the present, concluding this chapter of my story. This summer, I will be collaborating with CHAI, working under Dylan Hadfield-Menell and my new advisor to extend both *Inverse Reward Design* and *Whitelist Learning* (the latter being my proposal to CHAI; I plan to make a top-level post in the near future) [6].

Forwards

I sacrificed some of my tethering to the social web, working my way free of irrelevant external considerations, affirming to myself that I will look out for my interests. When I

first made that affirmation, I felt a palpable sense of *relief*. Truly, if we examine our lives with seriousness, what pressures and expectations bind us to arbitrary social scripts, to arbitrary identities - to arbitrary lives?

---

[1] My secret to being able to continuously soak up math is that I *enjoy it*. However, it wasn't immediately obvious that this would be the case, and only the intensity of my desire to step up actually got me to start studying. Only then, after occupying myself in earnest with those pages of Greek glyphs, did I realize that it's *fun*.

[2] This event marked my discovery of the mental movement detailed in *How to Dissolve It*; it has since paid further dividends in both novel ideas and clarity of thought.

[3] I've since updated away from this being true for humans in practice, but I felt it would be dishonest to edit my thought process after the fact.

Additionally, I did not fit any aspect of this story to the Sequences *post factum*; every reference was explicitly considered at the time (*e.g.*, remembering that specific post on how people don't usually give a serious effort *even when everything may be at stake*).

[4] I am so thankful to everyone who gave me advice. Summarizing for future readers:

- Speak in terms of the concrete problems in AI safety to avoid immediately getting pattern-matched.
- Frame project ideas (in part) with respect to their relevance to current ML systems.
- Explore all your funding options, including:
  - OpenPhilanthropy
  - Berkeley Existential Risk Initiative
  - Future of Life Institute
  - Paul Christiano's funding for independent alignment research

If you're navigating this situation, are interested in AI safety but want some direction, or are looking for a community to work with, please feel free to contact me.

[5] I'd like to emphasize that support for AI safety research is quickly becoming more mainstream in the professional AI community, and may soon become the majority position (if it is not already).

Even though ideas are best judged by their merits and not by their popular support, it can be emotionally important in these situations to remember that if you are concerned, you are *not* on the fringe. For example, 1,273 AI researchers have publicly declared their support for the Future of Life Institute's AI principles.

> A survey of AI researchers (Muller & Bostrom, 2014) finds that on average they expect a 50% chance of human-level AI by 2040 and 90% chance of human-level AI by 2075. On average, 75% believe that superintelligence (“machine intelligence that greatly surpasses the performance of every human in most professions”) will follow within thirty years of human-level AI. There are some reasons to worry about sampling bias based on *e.g.* people who take the idea of human-level AI seriously being more likely to respond (though see the attempts made to control for such in the survey) but taken seriously it suggests that most AI researchers

> think there’s a good chance this is something we’ll have to worry about within a generation or two.
>
> *AI Researchers on AI Risk (2015)*

[6] Objectives are subject to change.

Making a Difference Tempore: Insights from 'Reinforcement Learning: An Introduction'

> The safety of artificial intelligence applications involving reinforcement learning is a topic that deserves careful attention.

Foreword

Let's get down to business.

Reinforcement Learning

1: Introduction

2: Multi-armed Bandits

Bandit basics, including nonstationarity, the value of optimism for incentivizing exploration, and upper-confidence-bound action selection.

Some explore/exploit results are relevant to daily life - I highly recommend reading *Algorithms to Live By: The Computer Science of Human Decisions*.

3: Finite Markov Decision Processes

The framework.

4: Dynamic Programming

Policy evaluation, policy improvement, policy iteration, value iteration, generalized policy iteration. What a nomenclative nightmare.

5: Monte Carlo Methods

Prediction, control, and importance sampling.

Importance Sampling

After gathering data with our behavior policy b, we then want to approximate the value function for the target policy π. In off-policy methods, the policy we use to

gather the data is different from the one whose value v_π we're trying to learn; in other words, the distribution of states we sample is different. This gives us a skewed picture of v_π, so we must overcome this bias.

If b can take all of the actions that π can (*i.e.*, $\forall a, s : \pi(a|s) > 0 \implies b(a|s) > 0$), we can overcome by adjusting the return G_t of taking a series of actions $A_t, \dots, A_{T-1}$ using the importance-sampling ratio $\rho_{t:T-1} := \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$. This cleanly recovers v_π by the definition of expectation:

$$\begin{aligned}
\rho_{t:T-1} \mathbb{E}_b[G_t \mid S_t = s] &= \left(\prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}\right) \mathbb{E}_b[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= v_\pi(s).
\end{aligned}$$

Then after observing some set of returns (where $\{G_t\}_{t \in \mathcal{T}(s)}$ are the relevant returns for state s), we define the state's value as the average adjusted observed return

$$V(s) := \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$$

However, the $\rho_{t:T(t)-1}$'s can be arbitrarily large (suppose $\pi(A|S) = .5$ and $b(A|S) = 1 \times 10^{-10}$; $\frac{.5}{1 \times 10^{-10}} = .5 \times 10^{10}$), so the variance of this estimator can get pretty big. To get an estimator whose variance converges to 0, try

$$V(s) := \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Death and Discounting

Instead of viewing the discount factor γ as measuring how much we care about the future, think of it as encoding the probability we don't terminate on any given step. That is, we expect with $1 - \gamma$ probability to die on the next turn, so we discount rewards accordingly.

This intuition hugs pre-theoretic understanding much more closely; if you have just 80 years to live, you might dive that big blue sky. If, however, you imagine there's a non-

trivial chance that humanity can be more than just a flash in the pan, you probably care more about the far future.

6: Temporal-Difference Learning

The tabular triple threat: TD(0) *,* SARSA *, and* Q *-learning.*

Learning TD Learning

$$V(S_t) \leftarrow V(S_t) + \alpha [\underbrace{R_{t+1} + \gamma V(S_{t+1})}_{\text{TD target}} - V(S_t)]$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{TD error}}$$

SARSA

TD(0) ★ Q

- TD(0) is one-step bootstrapping of *state values* v_π. It helps you learn the value of a given policy, and is not used for control.
- SARSA is on-policy one-step bootstrapping of *action values* q_π using the quintuples (S, A, R, S', A').

As in all on-policy methods, we continually estimate q_π for the policy π, and at the same time change π toward greediness with respect to q_π.

- Q-learning is *off-policy* one-step bootstrapping of action values q_π.
    - You take an action using π and then use the maximal action value at the next state in your TD target term.

max**imization bias**

With great branching factors come great biases, and optimistic bias is problematic for Q-learning.

Refresher: the Q-learning update rule for state S_t, action A_t, and new state S_{t+1} is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)].$$

Suppose the rewards for all 100 actions at S_{t+1} are distributed according to $N(0, 1)$. All of these actions have a true (expected) value of 0, but the probability that none of their estimates are greater than .8 after 1 draw each is $\Phi(.8)^{100} \approx 4.6 \times 10^{-11}$. The more actions you have, the higher the probability is that at the maximum is just an outlier. See: regression toward the mean and mutual funds.

To deal with this, we toss another Q-learner into the mix; at any given update, one has their value updated and the other greedifies the policy. The double Q-learning scheme works because both Q-learners are unlikely to be biased in the same way. For some reason, this wasn't discovered until 2010.

7: n-step Bootstrapping

n *-step everything.*

8: Planning and Learning with Tabular Methods

Models, prioritized sweeping, expected vs. sample updates, MCTS, and rollout algorithms.

Roles Models Play

Distribution models include the full range of possible futures and their probabilities. For example, a distribution model for two fair coins: {HH: .25, HT: .5, TT: .25}.

Sample models just let you, well, sample. HH, HT, HT, HT, HH, TT, HH, HT... You could sample thousands of times and gain a high degree of confidence that the above distribution model is generating the data, but this wouldn't be your *only* hypothesis (granted, it might have the lowest Kolmogorov complexity).

Note that a distribution model also can function as a sample model.

9: On-policy Prediction with Approximation

We finally hit the good stuff*: value-function approximators and stochastic-/semi-gradient methods.*

10: On-policy Control with Approximation

11: Off-policy Methods with Approximation

The deadly triad *of function approximation, bootstrapping, and off-policy training converge to sow divergence and instability in your methods. Metal.*

12: Eligibility Traces

I was told to skip this chapter; the first half was my favorite part of the book.

TD(λ) uses a backwards-view eligibility trace to update features, which I find elegant.

Suppose you have some feature extraction function $\phi : S \to R^d$. Then you apply your TD update not only based on the features relevant at the current state, but also to the time-decaying traces of the features of previously-visited states. $\lambda \in [0, 1]$ sets how quickly this eligibility decay happens; $\lambda = 0$ recovers TD(0), while $\lambda = 1$ recovers Monte-Carlo return methods.

When I was a kid, there was a museum I loved going to - it had all kinds of wacky interactive devices for kids. One of them took sounds and "held them in the air" at a volume which slowly decreased as a function of how long ago the sound was made. State features are sounds, and volume is eligibility.

13: Policy Gradient Methods

The policy gradient theorem, REINFORCE, and actor-critic.

14: Psychology

Creating a partial mapping between reinforcement learning and psychology.

Mental, Territorial

There was a word I was looking for that "mental model" didn't quite seem to fit: "the model with respect to which we mentally simulate courses of action". CFAR's "inner sim" terminology didn't quite map either, as to me, that points to the *system-in-motion* more than *that-on-which-the-system-runs*. The literature dubs this a cognitive map.

> Individual place cells within the hippocampus correspond to separate locations in the environment with the sum of all cells contributing to a single map of an entire environment. The strength of the connections between the cells represents the distances between them in the actual environment. The same cells can be used

> for constructing several environments, though individual cells' relationships to each other may differ on a map by map basis. The possible involvement of place cells in cognitive mapping has been seen in a number of mammalian species, including rats and macaque monkeys. Additionally, in a study of rats by Manns and Eichenbaum, pyramidal cells from within the hippocampus were also involved in representing object location and object identity, indicating their involvement in the creation of cognitive maps.
>
> Wikipedia, *Cognitive map*

In light of my work on whitelisting, I've been thinking about why we're so "object-oriented" in our mental lives. An off-the-cuff hypothesis: to better integrate with the rest of our mental models, the visual system directly links up to our object schema. One such object is then recognized and engraved as a discrete "thing" in our map. Hence, we emotionally "know" that the world "really is" made up of objects, and isn't just a collection of particles.

> Most of the information used by people for the cognitive mapping of spaces is gathered through the visual channel (Lynch, 1960).

Lahav and Mioduser's research somewhat supports this idea, suggesting that blind people not only have lower-fidelity and more declarative (as opposed to procedural / interactive) cognitive maps, they're also less likely to provide object-to-object descriptions.

Epistemic status: I made a not-obviously-wrong hypothesis and found two pieces of corroborating evidence. If anyone who knows more about this wants to chime in, please do!

15: Neuroscience

The reward prediction error hypothesis, dopamine, neural actor-critic, hedonistic neurons, and addiction.

16: Applications and Case Studies

From checkers to checkmate.

17: Frontiers

Temporal abstraction, designing reward signals, and the future of reinforcement learning. In particular, the idea I had for having a whitelist-enabled agent predict expected object-level transitions is actually one of the frontiers: general value functions. Rad.

Pandora's AI Boxing

> The rapid pace of advances in AI has led to warnings that AI poses serious threats to our societies, even to humanity itself.

This chapter talks a fair amount about the serious challenges in AI alignment (not sure if you all have heard of that problem), which is heartening.

> As to safety, hazards possible with reinforcement learning are not completely different from those that have been managed successfully for related applications of optimization and control methods.

I'm not sure about that. Admittedly, I'm not as familiar with those solutions, but the challenge here seems to be of a qualitatively different caliber. Conditional on true AI's achievement, we'd want to have extremely high confidence that it pans out *before* we flip the switch. The authors acknowledge that

> it may be impossible for the agent to achieve the designer's goal no matter what its reward signal is.

I don't think it's impossible, but it's going to be extremely hard to get formal probabilistic guarantees. I mean, if you don't know an agent's rationality, you can't learn their utility function. If you do know their rationality but not their probability distribution over outcomes, you still can't learn their utility function.

This is just *one* of the many open problems in alignment theory. If that's not enough, there are always the unknown unknowns...

Final Thoughts

I read the "nearly-complete" draft of the second edition, available here. It was pretty good, but I did find most of the exercises either too easy or requiring considerable programming investment to set up the environment described. The former makes sense, as I've already taken a course on this, and I'm probably a bit above the introductory level.

Some graphs could have been more clearly designed, and some math in the proof of Linear TD(0)'s convergence (p. 206-207) is underspecified:

> In general, w_t will be reduced toward zero whenever A is positive definite,
>
> meaning $y^\top Ay > 0$ for real vector y.

An additional precondition: y can't be the zero vector.

> For a key matrix of this type, positive definiteness is assured if all of its columns sum to a nonnegative number.

Unless I totally missed what "this type" entails, this is false if taken at face value:

$$\begin{pmatrix} -2 & 5 \\ 5 & -2 \end{pmatrix}$$

has nonnegative column sums and is also indefinite, having eigenvalues of 3 and -7.

However, the claim *is* true in the subtle way they use it - they're *actually* showing that since the matrix is symmetric, real, and strictly diagonally dominant with positive diagonal entries, it's also positive definite. This could be made clearer.

In all, reading this book was definitely a positive experience.

Forwards

I'll be finishing *Analysis II* before moving on to Jaynes's *Probability Theory* in preparation for a class in the fall.

Dota 2

Recently, OpenAI made waves with their OpenAI Five Dota 2 bot. To REINFORCE what I just learned and solidified, I might make a post in the near future breaking down how *Five* differs from the *Alpha(Go) Zero* approach, quantifying my expectations for *The International* for calibration.

No Longer a Spectator

Four months and one week ago, I started my journey through the MIRI reading list. In those dark days, attempting a proof induced a stupor similar to that I encountered approaching a crush in grade school, my words and thoughts leaving me.

Six textbooks later and with a *little* effort, I'm able to prove things like the convergence of Monte Carlo integration to the Riemann integral, threading together lessons from *All of Statistics* and *Analysis I*; target in mind, words and thoughts remaining firmly by my side.

> The rapid pace of advances in artificial intelligence has led to warnings that artificial intelligence poses serious threats to our societies, even to humanity itself. The renowned scientist and artificial intelligence pioneer Herbert Simon anticipated the warnings we are hearing today...
>
> He spoke of the eternal conflict between the promise and perils of any new knowledge, reminding us of the Greek myths of Prometheus, the hero of modern science, who stole fire from the gods for the benefit of mankind, and Pandora, whose box could be opened by a small and innocent action to release untold perils on the world.
>
> Simon urged us to recognize that as designers of our future and *not mere spectators*, the decisions *we* make can tilt the scale in Prometheus' favor.

If you are interested in working together to learn MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg, if you're in a scale-tilting *mood - I would be more than happy to work with you. Messaging me may also net you an invitation to the MIRIx Discord server.*

Turning Up the Heat: Insights from Tao's 'Analysis II'

Foreword

It's been too long - a month and a half since my last review, and about three months since *Analysis I*. I've been immersed in my work for CHAI, but reality doesn't grade on a curve, and I want more mathematical firepower.

On the other hand, I've been cooking up something really special, so watch this space!

Analysis II

12: Metric Spaces

Metric spaces; completeness and compactness.

Proving Completeness

It sucks, and I hate it.

13: Continuous Functions on Metric Spaces

Generalized continuity, and how it interacts with the considerations introduced in the previous chapter. Also, a terrible introduction to topology.

There's a lot I wanted to say here about topology, but I don't think my understanding is good enough to break things down - I'll have to read an actual book on the subject.

14: Uniform Convergence

Pointwise and uniform convergence, the Weierstrass M *-test, and uniform approximation by polynomials.*

Breaking Point

Suppose we have some sequence of functions $f^{(n)} : [0,1] \to \mathbb{R}$, $f^{(n)}(x) := x^n$, which converge pointwise to the 1-indicator function $f : [0,1] \to \mathbb{R}$ (*i.e.*, $f(1) = 1$ and 0 otherwise). Clearly, each $f^{(n)}$ is (infinitely) differentiable; however, the limiting function

f isn't differentiable at all! Basically, pointwise convergence isn't at all strong enough to stop the limit from "snapping" the continuity of its constituent functions.

Progress

As in previous posts, I mark my progression by sharing a result derived without outside help.

Already proven: $\int_{-1}^{1}(1-x^2)^N\,dx \geq \frac{1}{\sqrt{N}}$.

Definition. Let $\epsilon > 0$ and $0 < \delta < 1$. A function $f : R \to R$ is said to be an (ϵ, δ)-*approximation to the identity* if it obeys the following three properties:

- f is compactly supported on $[-1, 1]$.
- f is continuous, and $\int_{-\infty}^{\infty} f = 1$.
- $|f(x)| \leq \epsilon$ for all $\delta \leq |x| \leq 1$.

Lemma: For every $\epsilon > 0$ and $0 < \delta < 1$, there exists an (ϵ, δ)-approximation to the identity which is a polynomial P on $[-1, 1]$.

Proof of Exercise 14.8.2(c). Suppose $c \in R, N \in N$; define $f(x) := c(1-x^2)^N$ for $x \in [-1, 1]$ and 0 otherwise. Clearly, f is compactly supported on $[-1, 1]$ and is continuous. We want to find c, N such that the second and third properties are satisfied. Since $(1-x^2)^N$ is non-negative on $[-1, 1]$, c must be positive, as f must integrate to 1. Therefore, f is non-negative.

We want to show that $|c(1-x^2)^N| \leq \epsilon$ for all $\delta \leq |x| \leq 1$. Since f is non-negative, we may simplify to $(1-x^2)^N \leq \epsilon$. Since the left-hand side is strictly monotone increasing on $[-1, -\delta]$ and strictly monotone decreasing on $[\delta, 1]$, we substitute $x = \delta$ without loss of generality. As $\epsilon > 0$, so we may take the reciprocal and multiply by ϵ, arriving at $\epsilon(1-\delta^2)^{-N} \geq c$.

We want $\int_{-\infty}^{\infty} f = 1$; as f is compactly supported on $[-1, 1]$, this is equivalent to $\int_{-1}^{1} f(x)\,dx = 1$. Using basic properties of the Riemann integral, we have $\int_{-1}^{1} (1 - x^2)^N\,dx = \frac{1}{c}$. Substituting in for c,

$$\epsilon^{-1}(1-\delta^2)^N \leq \frac{1}{\sqrt{N}} \leq \int_{-1}^{1} (1-x^2)^N\,dx,$$

with the second inequality already having been proven earlier. Note that although the first inequality is not always true, we can make it so: since ϵ is fixed and $1 - \delta^2 \in (0, 1)$, the left-hand side approaches 0 more quickly than $\frac{1}{\sqrt{N}}$ does. Therefore, we can make N as large as necessary; isolating ϵ,

$$\epsilon \geq (1-\delta^2)^N \sqrt{N}$$

$$\epsilon \geq \sqrt{N} > (1-\delta^2)^N \sqrt{N},$$

the second line being a consequence of $1 > (1 - \delta^2)^N$. Then set N to be any natural number such that this inequality is satisfied. Finally, we set $c = \frac{1}{\int_{-1}^{1}(1-x^2)^N\,dx}$. By construction, these values of c, N satisfy the second and third properties. □

Convoluted No Longer

Those looking for an excellent explanation of convolutions, look no further!

Weierstrass Approximation Theorem

Theorem. Suppose $f : [a, b] \to R$ is continuous and compactly supported on $[a, b]$. Then for every $\epsilon > 0$, there exists a polynomial P such that $||P - f||_\infty < \epsilon$.

In other words, any continuous, real-valued f on a finite interval can be approximated with arbitrary precision by polynomials.

Why I'm talking about this. On one hand, this result makes sense, especially after taking machine learning and seeing how polynomials can be contorted into basically whatever shape you want.

On the other hand, I find this theorem intensely beautiful. $\overline{\overline{P[a,b]} = C[a,b]}$'s proof was slowly constructed, much to the reader's benefit. I remember the very moment the proof sketch came to me, newly-installed gears whirring happily.

15: Power Series

Real analytic functions, Abel's theorem, exp *and* log*, complex numbers, and trigonometric functions.*

EXP

Cached thought from my CS undergrad: exponential functions always end up growing more quickly than polynomials, no matter the degree. Now, I finally have the gears to see why:

$$\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

exp has *all* the degrees, so no polynomial (of necessarily finite degree) could ever hope to compete! This also suggests why $\frac{d}{dx} e^x = e^x$.

Complex Exponentiation

You can multiply a number by itself some number of times.

[*nods*]

You can multiply a number by itself a negative number of times.

[Sure.]

You can multiply a number by itself an irrational number of times.

[OK, I understand limits.]

You can multiply a number by itself an imaginary number of times.

[Out. Now.]

Seriously, this one's weird (rather, it *seems* weird, but how can "how the world is" be "weird")?

Suppose we have some $c \in C$, where $c = a + bi$. Then $e^c = e^a e^{bi}$, so "all" we need to figure out is how to take an imaginary exponent. Brian Slesinsky has us covered.

Years before becoming involved with the rationalist community, Nate asks this question, and Qiaochu answers.

This isn't a coincidence, because nothing is ever a coincidence.

Or maybe it is a coincidence, because Qiaochu answered every question on StackExchange.

16: Fourier Series

Periodic functions, trigonometric polynomials, periodic convolutions, and the Fourier theorem.

17: Several Variable Differential Calculus

A beautiful unification of Linear Algebra and calculus: linear maps as derivatives of multivariate functions, partial and directional derivatives, Clairaut's theorem, contractions and fixed points, and the inverse and implicit function theorems.

Implicit Function Theorem

If you have a set of points in R^n, when do you know if it's secretly a function

$g : R^{n-1} \to R$? For functions $R \to R$, we can just use the geometric "vertical line test" to figure this out, but that's a bit harder when you only have an algebraic definition. Also, sometimes we can implicitly define a function locally by restricting its domain (even if no explicit form exists for the whole set).

Theorem. Let E be an open subset of R^n, let $f : E \to R$ be continuously differentiable,

and let $y = (y_1, \dots, y_n)$ be a point in E such that $f(y) = 0$ and $\frac{\partial f}{\partial x_n} \neq 0$. Then there exists

an open $U \subseteq R^{n-1}$ containing $(y_1, \dots, y_{n-1})$, an open $V \subseteq E$ containing y, and a function

$g : U \to R$ such that $g(y_1, \dots, y_{n-1}) = y_n$, and

$$\{ (x_1 , \dots , x_n) \in V : f (x_1 , \dots , x_n) = 0 \} =$$

$$\{ (x_1 , \dots , x_{n-1} , g (x_1 , \dots , x_{n-1})) : (x_1 , \dots , x_{n-1}) \in U \} .$$

So, I think what's really going on here is that we're using the derivative at this known zero to locally linearize the manifold we're operating on (similar to Newton's

approximation), which lets us have some neighborhood U in which we can derive an implicit function, even if we can't always write it out.

18: Lebesgue Measure

Outer measure; measurable sets and functions.

Tao lists desiderata for an ideal measure before deriving it. Imagine that.

19: Lebesgue Integration

Building up the Lebesgue integral, culminating with Fubini's theorem.

Conceptual Rotation

Suppose $\Omega \subseteq R^n$ is measurable, and let $f : \Omega \to [0, \infty]$ be a measurable, non-negative function. The Lebesgue integral of f is then defined as

$$\int_\Omega f := \sup \{ \int_\Omega s : s \text{ is simple and non-negative, and minorizes } f \} .$$

This hews closely to how we defined the *lower* Riemann integral in Chapter 11; however, we don't need the equivalent of the upper Riemann integral for the Lebesgue integral.

To see why, let's review why Riemann integrability demands the equality of the lower and upper Riemann integrals of a function g. Suppose that we integrate over $[0, 1]$, and that g is the indicator function for the rationals. As the rationals are dense in the reals, any interval $[a, b] \subseteq [0, 1]$ ($b > a$) contains rational numbers, no matter how much the interval shrinks! Therefore, the upper Riemann integral equals 1, while the lower equals 0 (for similar reasons). g *is* Lebesgue integrable; since it's 0 almost everywhere (as the rationals have 0 measure), its integral is 0.

This marks a fundamental shift in how we integrate. With the Riemann integral, we consider the lim sup and lim inf of increasingly-refined upper and lower Riemann sums - this is the *length* approach. In Lebesgue integration, however, we consider which $E \subseteq \Omega$ is responsible for each value y in the range (*i.e.*, $f^{-1}(y) = E$), multiplying y by the measure of E - this is *inversion*.

In a sense, the Lebesgue integral more cleanly strikes at the heart of what it *means* to integrate. Surely, Riemann integration was not far from the mark; however, if you

rotate the problem slightly in your mind, you will find a better, cleaner way of structuring your thinking.

Final Thoughts

Although Tao botches a few exercises and the section on topology, I'm a big fan of *Analysis I* and *II*. Do note, however, that *II* is far more difficult than *I* (not just in content, but in terms of the exercises). He generally provides relevant, appropriately-difficult problems, and is quite adept at helping the reader develop rigorous and intuitive understanding of the material.

Forwards

Next is Jaynes' *Probability Theory*.

Tips

- To avoid getting hung up in Chapter 17, this book should be read after a linear algebra text.
- Don't do exercise 17.6.3 - it's wrong.
- Deep understanding comes from sweating it out. Don't hide, don't wave away bothersome details - stay and explore. If you follow my strategy of quickly generating outlines - can you formally and precisely write out each step?

Verification

I completed every exercise in this book; in the second half, I started avoiding looking at the hints provided by problems until I'd already thought for a few minutes. Often, I'd solve the problem and then turn to the hint: "be careful when doing *X* - don't forget edge case *Y*; hint: use lemma *Z*"! A pit would form in my stomach as I prepared to locate my mistake and back-propagate where-I-should-have-looked, before realizing that I'd *already* taken care of that edge case using that lemma.

Why Bother?

One can argue that my time would be better spent picking up things as I work on problems in alignment. However, while I've made, uh, quite a bit of progress with impact measures this way, concept-shaped holes are impossible to notice. If there's some helpful information-theoretic way of viewing a problem that I'd only realize if I had *already taken* information theory, I'm out of luck.

Also, developing mathematical maturity brings with it a more rigorous thought process.

Fairness

There's a sense I get where even though I've made immense progress over the past few months, it still *might not be enough*. The standard isn't "am I doing impressive things for my reference class?", but rather the stricter "am I good enough to solve serious problems that might not get solved in time otherwise?". This is quite the standard, and even given my textbook and research progress (including the upcoming posts), I don't think I meet it.

In a way, this excites me. I welcome any advice for buckling down further and becoming yet stronger.

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also net you an invitation to the MIRIx Discord server.

On a related note: thank you to everyone who has helped me; in particular, TheMajor has been incredibly generous with their explanations and encouragement.

And My Axiom! Insights from 'Computability and Logic'

Foreword

Max Tegmark's *Our Mathematical Universe* briefly touches on a captivating, beautiful mystery:

> The arrows indicate the close relations between mathematical structures, formal systems and computations. The question mark suggests that these are all aspects of the same transcendent structure, whose nature we still haven't fully understood.

The profound results compiled by the *Computability and Logic* textbook may be the first step towards the answer.

Computability and Logic

> If this sentence is true, then Santa Claus is real.

As usual, I'll explain confusions I had and generally share observations. This book is on the MIRI reading list.

1 Enumerability

Coming back to this book, I'm amazed by some of my margin scribbles – expressions of wonderment and awe at what now strike me as little more than obvious facts ("relations

are sets of ordered pairs!").

2 Diagonalization

Exercise 2.13 (*Richard's paradox*) What (if anything) is wrong with following argument?

> The set of all finite strings of symbols from the alphabet, including the space, capital letters, and punctuation marks, is enumerable; and for definiteness let us use the specific enumeration of finite strings based on prime decomposition. Some strings amount to definitions in English of sets of positive integers and others do not; strike out the ones that do not, and we are left with an enumeration of all definitions in English of sets of positive integers, or, replacing each definition by the set it defines, and enumeration of all sets of positive integers that have definitions in English. Since some sets have more than one definition, there will be redundancies. Strike them out to obtain an irredundant enumeration of all sets of positive integers that have definitions in English. Now consider the set of positive integers defined by the condition that a positive integer n belongs to the set if and only if does not belong to the nth set in the irredundant enumeration just described.
>
> This set does not appear in that enumeration, so it cannot have a definition in English. Yet it does, and in fact we have just given such a definition.

My first time reading this exercise, I had plenty of objections. "Is not abusive to use English in place of a formal system? How do we even quantify the expressiveness of English, is that a thing?", and so on. Yet, returning with more context and experience, a few minutes thought revealed to me the crux: information and enumerability aren't just defined by what is present, but by what's *not*.[⋆]

Let's take a little detour. Consider the regular expression 1^*; its language is infinite, and certainly computable. We don't even need a Turing machine to recognize it; a strictly less-expressive finite state automaton would do just fine. And yet there are infinite subsets of this language which are not at all computable.

Consider some reasonable encoding $\langle M, w \rangle$ of a Turing machine M and input w. As we see later, we can enumerate all possible Turing machines and inputs (given that we first fix the alphabet, etc.). This means that we can number the encodings. Consider the halting set; that is, $\{\langle M, w \rangle \mid M \text{ halts on input } w\}$. Expressed in unary, the numbers of the encodings belonging to this set is a strict subset of the regular language 1^*, and yet is not computable (because the halting set is not negatively recursively semi-decidable; *i.e.*, we can't say a computation *won't* halt. Thus, its complement is not computable).

Do you see the problem now?

[⋆] Here, we should be careful with how we interpret "information". After all, coNP-complete problems are trivially *Cook* reducible to their NP-complete counterparts (*e.g.*, query the oracle and then negate the output), but many believe that there isn't a corresponding *Karp* reduction (where we do a polynomial amount of computation before querying the oracle and returning its answer). Since we aren't considering complexity but instead whether it's computable at all, complementation is fine.

3 Turing Computability

> Turing's thesis is that any effectively computable function is Turing computable, so that computation in the precise technical sense we have been developing coincides with effective computability in the intuitive sense.

On several occasions, the authors emphasize how the intuitive nature of "effective computability" renders futile any attempt to formalize the thesis. However, I'm rather interested in formalizing intuitive concepts and therefore wondered why this hasn't been attempted. Indeed, it seems that a recent thesis by Vinogradova conducts a category-theoretic formalization of the notion of abstract computability (although since I don't yet know category theory, I can't tell how related it is).

4 Uncomputability

5 Abacus Computability

6 Recursive Functions

> The part where you say “FLooPs” and give up on Turing-complete primitive recursion when the theorems don’t support it.
>
> Preface, *The Sequents*

Nate wrote:

> -Zero: The function that always returns 0.
>
> -Successor: The function that, given n, returns n + 1.
>
> -Identity: Functions that return something they are given.
>
> -Composition: The function that performs function composition.
>
> These four functions are called the "primitive recursive" functions, and some time is spent exploring what they can do. If we allow use of a fifth function:
>
> -Minimization: Given f, finds the arguments for which f returns 0.
>
> we get the "recursive" functions.

However, the book defines minimization like so:

$$\mathrm{Mn}[f](x_1, \ldots, x_n) = \begin{cases} y & \text{if } f(x_1, \ldots, x_n, y) = 0 \text{ and for all } t < y, \\ & f(x_1, \ldots, x_n, t) \text{ is defined and } \neq 0 \\ \text{undefined} & \text{if there is no such } y. \end{cases}$$

This confused me for days, and I didn't truly understand it until I came back several months later (*i.e.*, now). How in the world is it effectively computable if it isn't even

defined on all inputs?

Suppose I challenge you to give me a function that, given a number x, returns a larger number. The catch is that you aren't allowed to directly modify x – you can only use it to check whether your candidate solution is bigger. If you just use the bounded search provided by primitive recursion, for some valid inputs your function will be wrong. If you have to start from scratch, there's no finite number of exponentiations or tetrations or super-duper-tetrations that you can include which will work for all inputs. You have to be able to do unbounded search – for example, taking the successor until you get a larger number.

Depending on the domain, this isn't always total, either. If we're working with R^+ and I give you ∞, you'll increment forever. Similarly, your function won't be defined on input cat. The important part is that we've given an effective procedure for finding the solution whenever it exists and for valid inputs.

7 Recursive Sets and Relations

8 Equivalent Definitions of Computability

Coming to appreciate this bridge between math and computer science was one of my most profound experiences last year. My mind's eye began viewing the world differently. Goings-on came to be characterized not just as interactions of atomic "objects", but as the outgrowth of the evolution of some mathematical structure. As a vast and complex play of sorts, distorted by my mind and approximated in specific ways – some legible, others not.

Most of all, a tugging in the back of my mind intensified, continually reminding me just how ignorant I am about the nature of our world.

9 A Précis of First-Order Logic: Syntax

10 A Précis of First-Order Logic: Semantics

> In fact, if you take Euclid's first four postulates, there are many possible interpretations in which "straight line" takes on a multitude of meanings. This ability to disconnect the *intended* interpretation from the *available* interpretations is the bedrock of model theory. Model theory is the study of a*ll* interpretations of a theory, not just the ones that the original author intended.
>
> Of course, model theory isn't really about finding surprising new interpretations — it's much more general than that. It's about exploring the breadth of interpretations that a given theory makes available. It's about discerning properties that hold in all possible interpretations of a theory. It's about discovering how well (or poorly) a given theory constrains its interpretations. It's a toolset used to discuss interpretations in general.
>
> At its core, model theory is the study of what a mathematical theory actually says, when you strip the intent from the symbols.

I can't emphasize enough how helpful Nate's *Mental Context for Model Theory* was; the mental motions behind model theory are a major factor in my excitement for studying

more logic.

11 The Undecidability of First-Order Logic

12 Models

Coming out of linear algebra with a "isomorphism ?= bijection" confusion, the treatment in this chapter laid the conceptual foundation for my later understanding homomorphisms. That is, a key part of the "meaning" of mathematical objects lies not just in their number, but in how they relate to one another.

This chapter is also great for disassociating baggage we might naïvely assign to words on the page, underlining the role of syntax as pointers to mathematical objects.

13 The Existence of Models

14 Proofs and Completeness

15 Arithmetization

16 Representability of Recursive Functions

I confess that upon wading back into the thicket of logical notation and terminology, I found myself lost. I was frustrated at how quickly I'd apparently forgotten everything I'd worked so hard for. After simmering down, I went back through a couple chapters, and found myself rather bored by how easy it was. I hadn't forgotten much at all – not beyond the details, anyways. I don't know whether that counts as "truly a part of me", but I don't think it's reasonable to expect myself to memorize everything, especially on the first go.

17 Indefinability, Undecidability, Incompleteness

> Perhaps the most important implication of the [first] incompleteness theorem is what it says about the notions of *truth* (in the standard interpretation) and *provability* (in a formal system): *that they are in no sense the same*.

Indeed, the notion of provability can be subtly different from our mental processes for judging the truth of a proposition; within the confines of a formal system, provability doesn't just tell us about the proposition in question, but also about the characteristics of that system. This *must* be kept in mind.

> When Godel's theorem first appeared, with its more general conclusion that a mathematical system may contain certain propositions that are undecidable within that system, it seems to have been a great psychological blow to logicians, who saw it at first as a devastating obstacle to what they were trying to achieve. Yet a moment's thought shows us that many quite simple questions are undecidable by deductive logic. There are situations in which one can prove that a certain property must exist in a finite set, even though it is impossible to exhibit any member of the set that has that property. For example, two persons are the sole witnesses to an event; they give

> opposite testimony about it and then both die. Then we know that one of them was lying, but it is impossible to determine which one.
>
> In this example, the undecidability is not an inherent property of the proposition or the event; it signifies only the incompleteness of our own *information*. But this is equally true of abstract mathematical systems; when a proposition is undecidable in such a system, that means only that its axioms do not provide enough information to decide it. But new axioms... might supply the missing information and make the proposition decidable after all.
>
> In the future, as science becomes more and more oriented to thinking in terms of information content, Godel's results will be seen as more of a platitude than a paradox. Indeed, from our viewpoint "undecidability" merely signifies that a problem is one that calls for *inference* rather than deduction. Probability theory as extended logic is designed specifically for such problems.
>
> ~ E.T. Jaynes, *Probability Theory*

18 The Unprovability Of Consistency

19 Normal Forms

20-27 [Skipped]

I found these advanced topics rather boring; the most important was likely provability logic, but I intend to study that separately in the future anyways.

Final Thoughts

- Nate Soares already covered this; unlike him, I didn't quite find it to be a breeze, although it certainly isn't the hardest material I covered last year.
- I'm surprised the authors didn't include the thematically appropriate recursion theorem, which states that a Turing machine can use its source code in its own computation without any kind of infinite regress (see: quines). This theorem allows particularly elegant proofs of the undecidability of the halting problem, and more generally, of Rice's theorem.

I really liked this book. In the chapters I completed, I did all of the exercises – they seemed to be of appropriate difficulty, and I generally didn't require help.

I've already completed *Understanding Machine Learning*, the first five chapters of *Probability Theory*, and much of two books on confrontational complexity. I'm working through a rather hefty abstract algebra tome and intend to go through two calculus books before an ordinary differential equations text. The latter material is more recreational, as I intend to start learning physics. This project will probably be much slower, but I'm really looking forward to it.

Forwards

> Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.

> ~ Banach

I don't think I'm a great mathematician yet by any means, but as my studies continue, I can't shake a growing sense of structure. I'm trying to broaden my toolbox as much as possible, studying areas of math which had seemed distant and unrelated. Yet the more I learn, the more my mental buckets collapse and merge.

And to think that I had once suspected the Void of melodrama.

> If for many years you practice the techniques and submit yourself to strict constraints, it may be that you will glimpse the center. Then you will see how all techniques are one technique, and you will move correctly without feeling constrained. Musashi wrote: “When you appreciate the power of nature, knowing the rhythm of any situation, you will be able to hit the enemy naturally and strike naturally. All this is the Way of the Void.”

If you are interested in working with me or others to learn MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Message me for an invitation to the MIRIx Discord server.

Judgment Day: Insights from 'Judgment in Managerial Decision Making'

I found this book through the CFAR reading list. Some content was previously posted on my shortform feed.

Foreword

The more broadly I read and learn, the more I bump into implicit self-conceptions and self-boundaries. I was slightly averse to learning from a manager-oriented textbook because I'm not a manager, but also because I... didn't see myself as the kind of person who could learn about a "business"-y context? I also didn't see myself as the kind of person who could read and do math, which now seems ridiculous.

Although the read was fast and easy and often familiar, I unearthed a few gems.

Judgment in Managerial Decision Making

A tip of dubious ethicality for lazy min-maxers:

> Managers give more weight to performance during the three months prior to [a performance] evaluation than to the previous nine months... because it is more available in memory.

Unified explanation of biases

The authors group biases as stemming from the Big Three of availability, representativeness, and confirmation. The model I took away relies on a mechanism somewhat similar to attention in neural networks: due to how the brain performs time-limited search, more salient/recent memories get prioritized for recall.

The availability heuristic goes wrong when our saliency-weighted perceptions of the frequency of events is a biased estimator of the real frequency, when we happen to be extrapolating off of a very small sample size, or when our memory structure makes recalling some kinds of things harder (e.g. words starting with 'a' versus words whose third letter is 'a'). Concepts get inappropriately activated in our mind, and we therefore reason incorrectly. Attention also explains anchoring: you can more readily bring to mind things related to your anchor due to salience.

The representativeness heuristic can be understood as highly salient concept-activations inappropriately dominating our reasoning. We then ignore e.g. base rates, sample size, statistical phenomena (regression to the mean), and the conjunctive

burden of propositions. Consider how neural network activations could explain the following:

> Intuitively, thinking of Linda as a feminist bank teller "feels" more correct than thinking of her as only a bank teller.

The case for confirmation bias seems to be a little more involved. We had evolutionary pressure to win arguments, which might mean our cognitive search aims to *find* supportive arguments and *avoid* even subconsciously signalling that we are aware of the existence of counterarguments. This means that those supportive arguments *feel* salient, and we (perhaps by "design") get to feel unbiased - we aren't consciously discarding evidence, we're just following our normal search/reasoning process! This is what our search algorithm feels like from the inside.

Making heads and tails of probabilistic reasoning

In *Subjective Probability: A Judgment of Representativeness*, Kahneman and Tversky hypothesize that

> since sample size does not represent any property of the population, it is expected to have little or no effect on judgment of likelihood,

which lines up with the above attention/activation model. Anyways, participants judged that the sequence of birth sexes GBGBBG is *more likely* than BGBBBB (obviously, they have equal probability). K&T chalk this up to the first sequence seeming more "representative" of a "random" process. If you're considering whether the set of all sequences which "look like" the former is more likely than the set of sequences resembling the latter, then this answer could be correct.

However, checking the original paper, this was controlled for; K&T emphasized that the exact order of births was as described. They go on:

> One may wonder whether [subjects] do not simply ignore order information, and answer the question by evaluating the frequency of families of five boys and one girl relative to that of families of three boys and three girls. However, when we asked the same [subjects] to estimate the frequency of the sequence BBBGGG, they viewed it as significantly less likely than GBBGBG ($p < .01$), presumably because the former appears less random. Order information, therefore, is not simply ignored.

Share your unique information in groups

Groups are good because they pool knowledge and expertise. However, studies show that by default, shared knowledge is much more likely to be discussed than unshared knowledge, which can significantly worsen decision-making. The authors give the example of a group initially disfavoring a student council candidate. One person is privately aware of crucial positive information about the candidate, and groups in

which all members knew the info were likely to favor the candidate. The information wasn't usually shared, and the candidate was passed over.

> [Ameliorative] strategies include forewarning the group in advance of the unique knowledge of different members and identifing expertise in the group before the discussion begins.

Open subaccounts for savings

You can avoid psychological annoyances throughout the year (tickets, unanticipated fees, etc.) *and* counteract the budget-planning fallacy by, at the beginning of each year, allocating money to goal-specific subaccounts. Then, you can forget about it during the year, and (perhaps) donate the remainder to an effective charity.

Be more risk-neutral

> Paul Samuelson... offered a colleague a coin-toss gamble. If the colleague won the coin toss, he would receive $200, but if he lost, he would lose $100. Samuelson was offering his colleague a positive expected value with risk. The colleague, being risk-averse, refused the single bet, but said that he would be happy to toss the coin 100 times! The colleague understood that the bet had a positive expected value and that across lots of bets, the odds virtually guaranteed a profit. Yet with only one trial, he had a 50% chance of regretting taking the bet.
>
> Notably, Samuelson's colleague doubtless faced many gambles in life… He would have fared better in the long run by maximizing his expected value on each decision... all of us encounter such "small gambles" in life, and we should try to follow the same strategy. **Risk aversion is likely to tempt us to turn down each individual opportunity for gain. Yet the aggregated risk of all of the positive expected value gambles that we come across would eventually become infinitesimal, and potential profit quite large.**

Biological explanation for hedonic treadmill?

> The striking aspect about framing and reference-point effects is that they suggest the presence of underlying mental processes that are *more* complicated than a rational decision-maker would employ. Rational decision-makers would simply seek to maximize the expected value of their choices. Whether these outcomes represent gains or losses would be irrelevant, and consideration of the outcome relative to the status quo would be a superfluous consideration. instead, we adjust to the status quo, and then think of the changes from that point as gains or losses.
>
> Rayo and Becker (2007) present a persuasive explanation for why evolution programmed us with extra machinery that impairs our decisions. According to their explanation, our reliance on frames and reference points to assess outcomes is an elegant solution to a problematic biological constraint. The constraint is that our "subjective utility scale" – our ability to experience pleasure and pain – is not infinitely sensitive. Was Bill Gates's 50th billion dollars as satisfying as his first? certainly not. the limited sensitivity of our subjective utility scale is precisely the reason why we experience declining marginal utility for both gains and losses…

> Given this biological constraint on the sensitivity of our subjective utility scale, we need to readjust our reference point by getting used to what we've got and then taking it for granted. If we didn't adjust our reference point, we could quickly hit the maximum of our utility scale, and realize that nothing we could ever do would make us happier. That would effectively kill our motivation to work harder, become richer, and achieve more. In reality, of course, what happens is that we get used to our current level of wealth, status, and achievement, and are then motivated to seek more, believing that it will make us happier.
>
> The irony of this motivational system is that for it to keep working, **we have to habituate to our new condition but not anticipate this habituation.** Evidence does indeed confirm that people adjust to both positive and negative changes in circumstances with surprising speed, and then promptly forget that they did so. Thus, we find ourselves on a hedonic treadmill in which we strive for an imagined happiness that forever slips out of our grasp, beckoning us onward.

Negotiation tips

Chapters 9 and 10 contain a wealth of (seemingly) good negotiation advice. Being a good negotiator and mediator seems like an important generalist life skill.

- Before negotiation, consider all relevant issues and their importance to you and then to your partner. Try to spot places where you can make efficient positive-sum bargains. Figure out your next-best alternative if a deal cannot be struck, and anticipate what your partner's will be as well.
- Find the intent generating their stated position. Maybe your boss states they don't want you installing a standing desk, but they're secretly worried it'll lead to a slippery slope of employees installing increasingly distracting accessories. If you can find this out, you can offer your support in preventing a slippery slope, instead of trying to push on the more difficult all-or-nothing position.
- Negotiate multiple issues simultaneously. You're better able to find positive-sum agreements when considering multiple axes at once. Also, you'll avoid making them compromise too hard and too early on things which aren't important to you, which can make the later part of negotiation less fruitful for you.

There were a lot more helpful takeaways, and I plan on rereading Ch. 9 before conducting any important negotiations.

Forwards

This book was a little slow at times, both because of excessive preamble/signposting, and my already being familiar with much of the literature. Still, I'm glad I read it.

Hello again

It's been a long while since my last review. After injuring myself last summer, I wasn't able to type reliably until early this summer. This derailed the positive feedback loop I had around "learn math" -> "write about what I learned" -> "savor karma". Protect your feedback loops.

I run into fewer basic confusions than when I was just starting at math, so I generally have less to talk about. This means I'll be changing the style of any upcoming reviews, instead focusing on deeply explaining the things I found coolest.

Since January, I've read *Visual Group Theory, Understanding Machine Learning, Computational Complexity: A Conceptual Perspective, Introduction to the Theory of Computation, An Illustrated Theory of Numbers*, most of Tadellis' *Game Theory*, the beginning of *Multiagent Systems*, parts of several graph theory textbooks, and I'm going through Munkres' *Topology* right now. I've gotten through the first fifth of the first Feynman lectures, which has given me an unbelievable amount of mileage for generally reasoning about physics.

My "plan" is to keep learning math until the low graduate level (I still need to at least do complex analysis, topology, field / ring theory, ODEs/PDEs, and something to shore up my atrocious trig skills, and probably more)[1], and then branch off into physics + a "softer" science (anything from microecon to psychology).

New year, new decade

In the new year, I'm going to focus hard on raising the level of my cognitive game.

Reading the Sequences qualitatively levelled me up, and I want to do that again. My thought processes are still insufficiently transparent: I need to flag motivated reasoning more often. I still fall prey to the planning fallacy (but somewhat less than two years ago). I don't notice my confusion *nearly* as often as I should.

Not noticing confusion often has a cost measured in hours (or more). Let me give you an example. Last night, I went to speak with Sen. Amy Klobuchar about effective altruism. It was my understanding that the event would be a meet-and-greet. I planned to query her interest in e.g. setting up a granting agency disbursing funds based on scientific evidence of high impact, with the details to be worked out in conjunction with relevant professionals in EA and the government.

While I was waiting for her to arrive, I noticed that people were writing on paper and handing it to other people. I rounded this off as commit-to-caucus cards, which, if I *actually thought* about it, makes no sense – you keep your commit-to-caucus card. They were, in fact, providing written questions, some of which Sen. Klobuchar would later answer. If I had just noticed this, I could have written a question and then left, saving myself two hours.

The list of things I've noticed I failed to notice in the last month is surprisingly long. I don't think I'm bad at this in a relative sense – just in an absolute sense.

This new year, I'm going to become a less oblivious, less stupid, and less wrong person.

---

1. I also still want to learn Bayes nets, category theory, get a much deeper understanding of probability theory, provability logic, and decision theory. ↩

Continuous Improvement: Insights from 'Topology'

Foreword

Sometimes you really like someone, but you can't for the life of you understand why. By all means, you should have tired of them long ago, but you keep coming back for more. Welcome, my friend, to *Topology*.

This book is a good one, but boy was it *slow* (349 pages at ~30 minutes a page, on average). I just kept coming back, and I was slowly rewarded each time I did.

Note: sil ver already reviewed *Topology*.

Topology

Topology is about what it means for things to be "close" in a very abstract and general sense. Rather than taking on the monstrous task of intuitively explaining topology without math, I'm just going to talk about random things from the book and (literally) illustrate concepts which were at first confusing.

Compactness = wonderful kind of mathematical "smallness"

> Compact means small. It is a peculiar kind of small, but at its heart, compactness is a precise way of being small in the mathematical world. The smallness is peculiar because, as in the example of the open and closed intervals $(0, 1)$ and $[0, 1]$, a set can be made “smaller” (that is, compact) by adding points to it, and it can be made “larger” (non-compact) by taking points away.
>
> As a notion of smallness, then, compactness is a bit fraught. It’s a bit unsettling to say that a set can be “smaller” than a set that lies entirely inside it! But I think smallness is a valuable way to see compactness. A set that is compact may be large in area and complicated, but the fact that it is compact means we can interact with it in a finite way using open sets, the building blocks of topology.
>
> *What Does Compactness Really Mean?*

Minimum description length says that an explanation is big if its shortest computational specification is long. You can have a simple explanation of a very long list of things or of a large universe, and extremely complicated explanations of things easily expressed in natural language (God's source code would be *a lot* longer than Maxwell's equations).

VC dimension says a class of hypotheses is hard to learn if it has lots of predictive degrees of freedom. You can have an infinite class of hypotheses which is really easy to

learn because it has low VC dimension (thresholding functions at value θ), and a finite class which is really hard to learn because it has high VC dimension (all C programs less than 1 million characters).

Compactness says that a topological space is big if it has a covering of open sets that can't be trimmed down to a finite subcollection which still covers the whole space. You can have an uncountable compact space ($[0, 1]$ under the standard topology, or even a Cantor space), and a countable space which isn't compact (Q under the standard topology; note that all countable topological spaces have to at least be Lindelof).

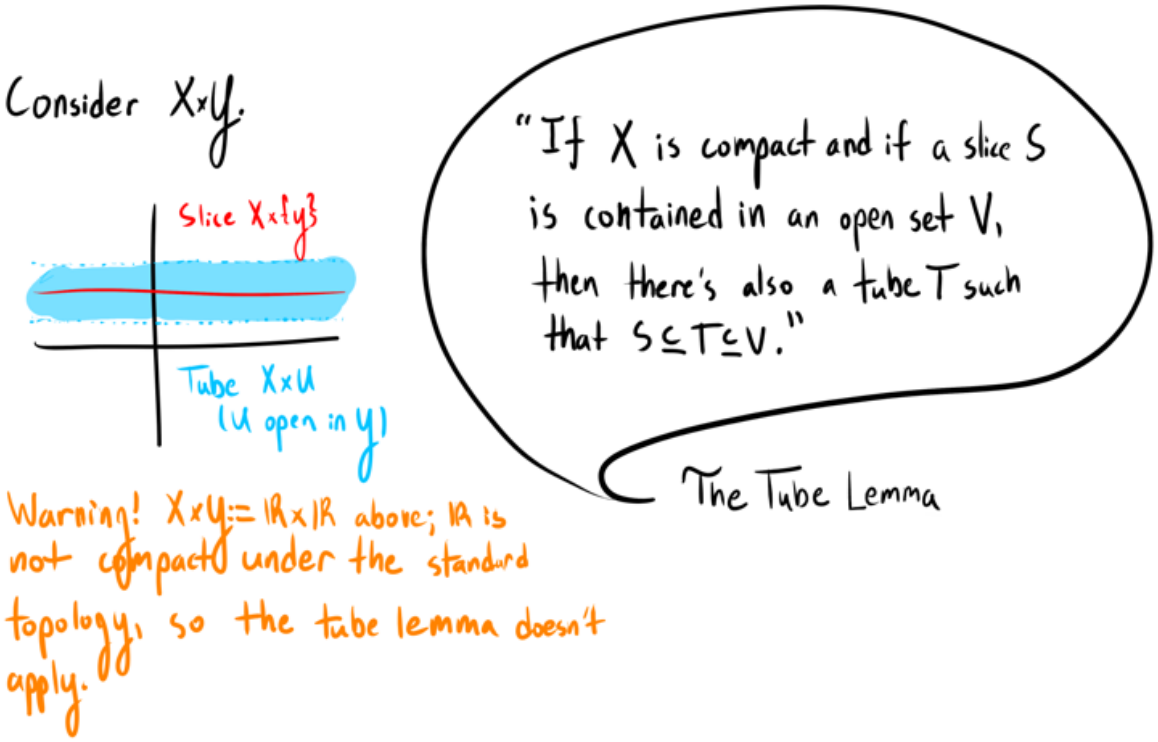
Compactness is not always inherited by open subspaces

At first, I was confused why *open* subspaces Y of compact X don't have to be compact (if Y is closed, it does have to be compact). But compactness requires *all* open coverings of Y to have a finite subcover. Meaning, you can't just give it X's finite cover intersect the subspace, because the finite subcover has to be a subcollection of Y's covering.

Getting closure

Theorem: If X *is compact, show that the projection* $\pi_2 : X \times Y \to Y$ *is closed.*

I was confused why we needed compactness. Essentially, I didn't understand the tube lemma.



Now let's prove the theorem. Suppose C is closed in $X \times Y$. We want to show $f(C)$ is also closed. Take $y \notin \pi(C)$. $(X \times Y) - C$ is an open set of the domain containing the slice $X \times \{y\}$. Since X is compact, apply the tube lemma to get a tube $X \times U$. The projection of this tube is both open (because U is open in Y) and disjoint from $\pi(C)$ (because the tube is contained in $(X \times Y) - C$). Thus, all $y \notin \pi(C)$ have an open neighborhood disjoint from $\pi(C)$, so $\pi(C)$ must be closed.

Let X *be a locally compact space. If* $f : X \to Y$ *is continuous, does it follow that* $f(X)$ *is locally compact? What if* f *is both continuous and open?*

It has to be both continuous and open; the reason I got confused here was it seemed like continuity should be enough. It *was* plain to me how to prove it given f open, but this SE post has a good counterexample for just f continuous.

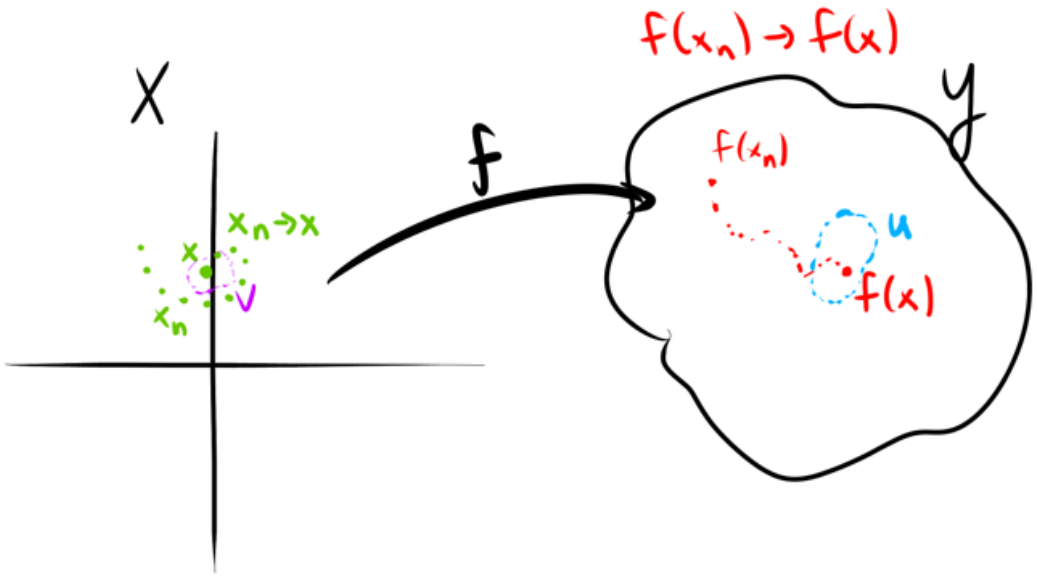
Multivariate continuity

How come you can have discontinuous multivariate functions which are continuous in each variable? What *is* continuity, with a product space as your domain? To simplify matters, let's consider two metric spaces X, Y.

One definition of continuity uses open sets – $f : X \to Y$ is continuous at x if, for every open neighborhood U of f(x), there exists an open neighborhood V of x such that $f(V) \subseteq U$.

Another definition uses topological convergence. $f : X \to Y$ is continuous at x if, for every sequence $x_n \to x$, $f(x_n) \to f(x)$.

These definitions are equivalent. The latter lets us think about how different winding paths you can take in a domain always must topologically converge to the same thing in the co-domain.



Continuity in the variables says that paths along the axes converge in the right way. But for continuity overall, we need *all* paths to converge in the right way. Directional continuity when the domain is R is a special case of this: continuity from below and from above if and only if continuity for all sequences converging topologically to x.

You only lift once

Suppose $p : C \to Y$ is a covering map. One way of understanding lifts in algebraic topology is that, for some path $f : X \to Y$, the lift $\tilde{f} : X \to C$ is the unique path in the covering space C corresponding to $f = p \circ \tilde{f}$.

EXAMPLE 1. Consider the covering $p : \mathbb{R} \to S^1$ of Theorem 53.1. The path $f : [0, 1] \to S^1$ beginning at $b_0 = (1, 0)$ given by $f(s) = (\cos \pi s, \sin \pi s)$ lifts to the path $\tilde{f}(s) = s/2$ beginning at 0 and ending at $\frac{1}{2}$. The path $g(s) = (\cos \pi s, -\sin \pi s)$ lifts to the path $\tilde{g}(s) = -s/2$ beginning at 0 and ending at $-\frac{1}{2}$. The path $h(s) = (\cos 4\pi s, \sin 4\pi s)$ lifts to the path $\tilde{h}(s) = 2s$ beginning at 0 and ending at 2. Intuitively, h wraps the interval $[0, 1]$ around the circle twice; this is reflected in the fact that the lifted path $\tilde{h}$ begins at zero and ends at the number 2. These paths are pictured in Figure 54.1.

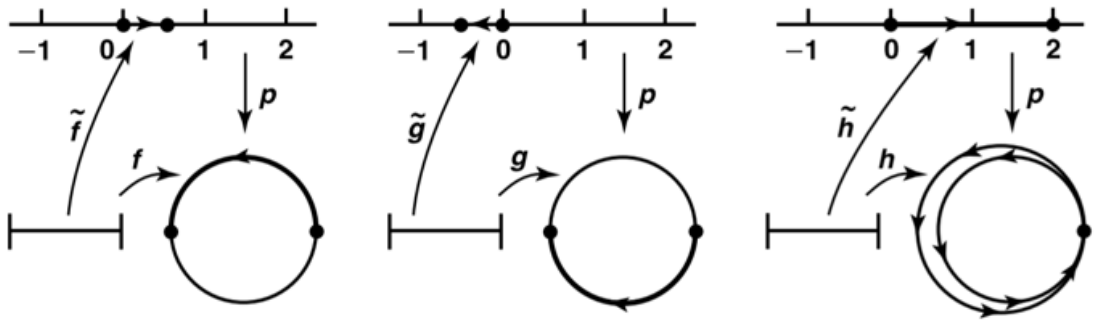
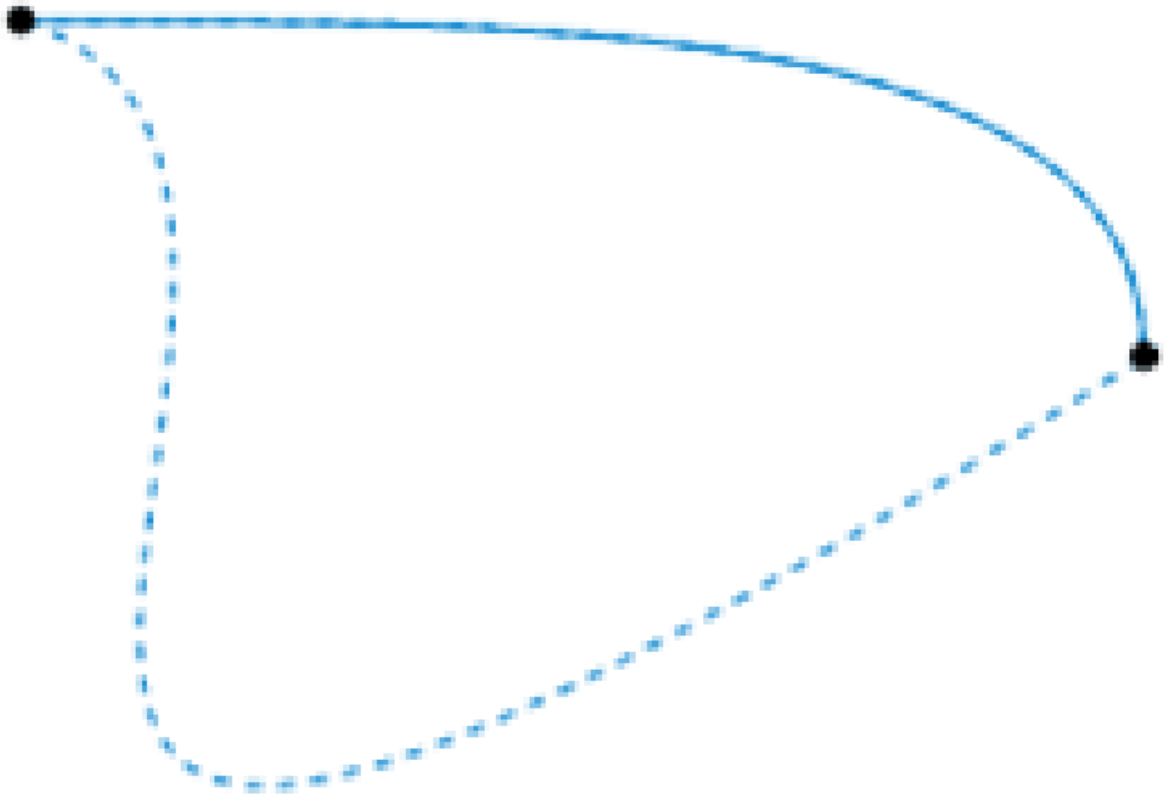


Figure 54.1

Once you fix the initial point, the lift corresponds to the unique path in the covering space which produces f. It's just helping you find the corresponding path in the lifted up covering space!

Homotopy



This concept yields amazing insight into such profound topics as the deeper nature of jump rope. Under the standard subspace topology of R^3, consider the space swept out by a rope held at fixed endpoints and tautness. All paths between the endpoints are path homotopic! You can think about movements of the rope (either clockwise or counterclockwise) as homotopies in this space.

Miscellaneous

- I stopped at about section 56 because I was getting diminishing returns. By this point, I felt like I had a solid understanding of point-set topology, and look forward to more thoroughly covering algebraic topology in the future.
- One-point compactifications feel like an important thing to grasp, and they're fun to play around with mentally. I skipped Stone-Cech compactification.
- Completeness in metric spaces means that Cauchy sequences converge topologically; in other words, nothing can "escape" from the space. I remember having problems with this (and with thinking about non-Hausdorff spaces) back when I was learning analysis. Things feel a lot better now.

Verdict

Topology can be dry, but it's exceedingly well-written and clear. I tried for quite a while to find a better topology book, but I didn't.

Forwards

Finally getting around to topology was such a good decision. For exercise solutions, see both MathOverflow and this site.

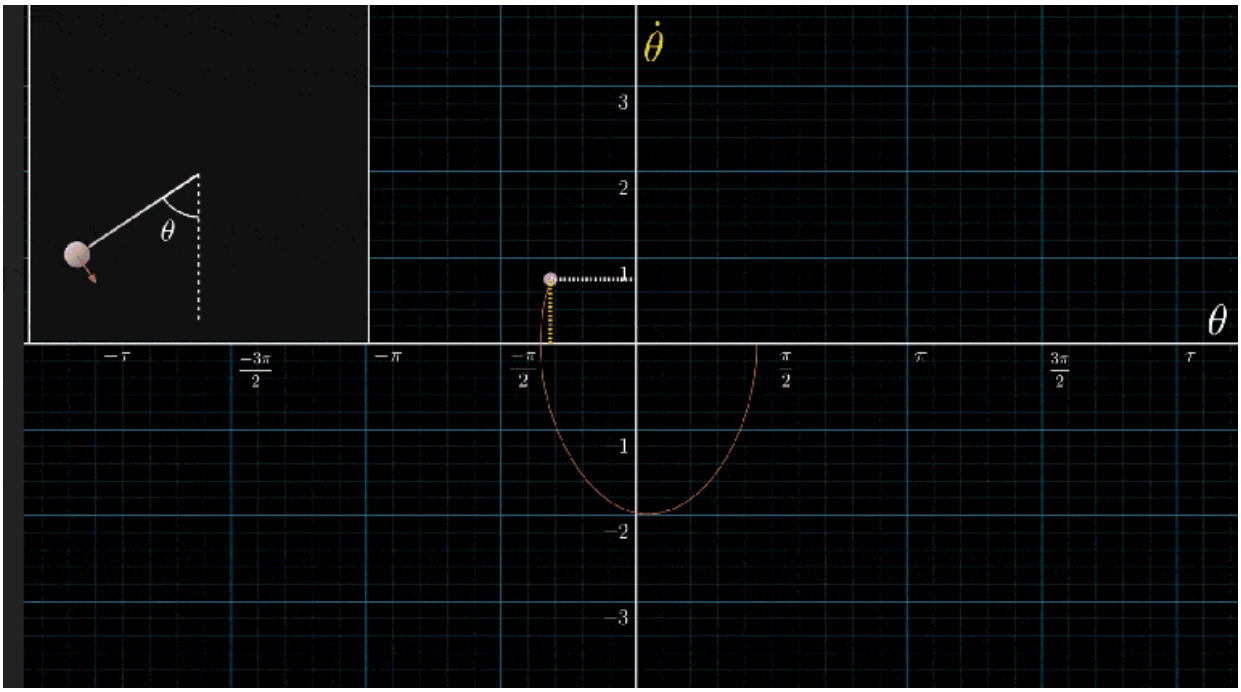
Some things change how you look at math, help you notice subtleties and shades and immediately grasp certain facets of new mathematical objects. Topology is one of these things, as is abstract algebra. Learning that an object is a group, or finitely generated, or isomorphic to a more familiar structure gives me an immediate head start. Similarly, learning that spaces are homeomorphic, or compact, or second-countable is *such* a boost.

What was I even *doing* with my life before I knew about homeomorphisms?

ODE to Joy: Insights from 'A First Course in Ordinary Differential Equations'

Foreword

Sometimes, it's easier to say how things change than to say how things are.



From [3Blue1Brown: Differential Equations]

When you write down a differential equation, you're specifying constraints and information about e.g. how to model something in the world. This gives you a family of solutions, from which you can pick out any function you like, depending on details of the problem at hand.

Today, I finished the bulk of Logan's *A First Course in Ordinary Differential Equations*, which is easily the best ODE book I came across.

A First Course in Ordinary Differential Equations

As usual, I'll just talk about random cool things from the book.

Bee Movie

In the summer of 2018 at a MIRI-CHAI intern workshop, I witnessed a fascinating debate: what mathematical function represents the *movie time* elapsed in videos like The Entire Bee Movie but every time it says bee it speeds up by 15%? That is, what mapping $t \mapsto x(t)$ converts the viewer timestamp to the movie timestamp for this video?

I don't remember their conclusion, but it's simple enough to answer. Suppose $f(t)$ counts how many times a character has said the word "bee" by timestamp t in the movie. Since the viewing speed itself increases exponentially with f, we have $x'(t) = 1.15^{f(x(t))}$. Furthermore, since the video starts at the beginning of the movie, we have the initial condition $x(0) = 0$.

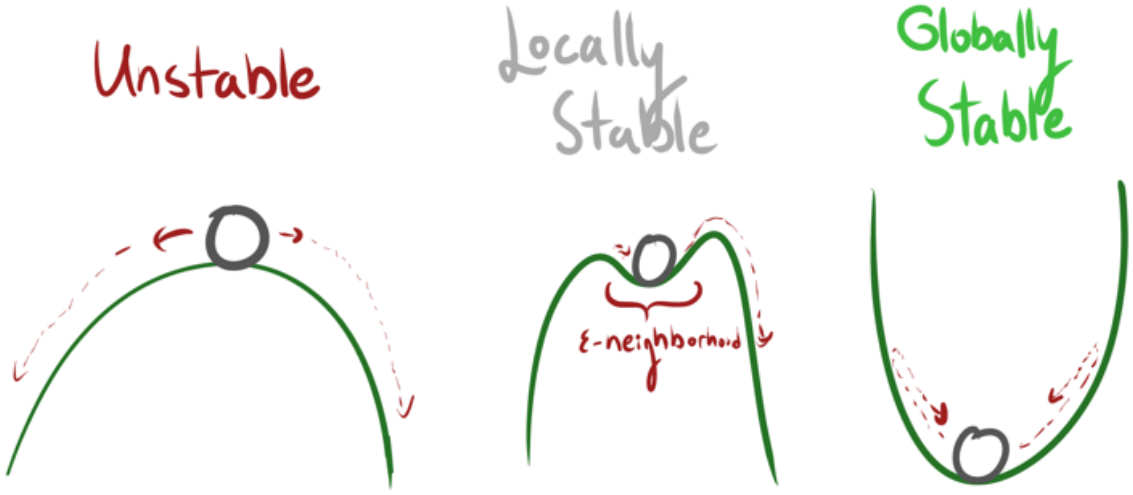
This problem cannot be cleanly solved analytically (because f is discontinuous and obviously lacking a clean closed form), but *is* expressed by a beautiful and simple differential equation.

Gears-level models?

Differential equations help us explain and model phenomena, often giving us insight into causal factors: for a trivial example, a population might grow more quickly *because* that population is larger.

Equilibria and stability theory

This material gave me a great conceptual framework for thinking about stability. Here are some good handles:



Let's think about rocks and hills. *Unstable* equilibria have the rock rolling away forever lost, no matter how lightly the rock is nudged, while *locally stable* equilibria have some level of tolerance within which they'll settle back down. For a *globally stable* equilibrium, no matter how hard the perturbation, the rock comes rolling back down the parabola.

Resonance

> A familiar example is a playground swing, which acts as a pendulum. Pushing a person in a swing in time with the natural interval of the swing (its resonant frequency) makes the swing go higher and higher (maximum amplitude), while attempts to push the swing at a faster or slower tempo produce smaller arcs. This is because the energy the swing absorbs is maximized when the pushes match the swing's natural oscillations. ~ Wikipedia

And that's also how the Tacoma bridge collapsed in 1940. The second-order differential equations underlying this allow us to solve for the forcing function which could induce catastrophic resonance.

Also note that there is only at most *one* resonant frequency of any given system, because even lower octaves of the natural frequency would provide destructive interference a good amount of the time.

Random notes

- This book gave me great chance to review my calculus, from integration by parts to the deeper meaning of Taylor's theorem: that for many functions, you can recover all of the global information from the local information, in the form of derivatives. I don't fully understand why this doesn't work for some functions which are infinitely differentiable (like $\log x$), but apparently this becomes clearer after some complex analysis.

- Bifurcation diagrams allow us to model the behavior, birth, and destruction of equilibria as we vary parameters in the differential equation. I'm looking forward to learning more about bifurcation theory. In this video, Veratasium highlights stunning patterns behind the bifurcation diagrams of single-humped functions.

Forwards

I supplemented my understanding with the first two chapters of Strogatz's *Nonlinear Dynamics And Chaos*. I might come back for more of the latter at a later date; I'm feeling like moving on and I think it's important to follow that feeling.

A Kernel of Truth: Insights from 'A Friendly Approach to Functional Analysis'

Foreword

What is functional analysis? A satisfactory answer requires going back to where it all started.

> "All are present; the meeting convenes," intoned Fredholm. Intent were the gathered faces, their thoughts fixed on their students. "*What do we know of their weaknesses?*"
>
> Hilbert leaned back, torch's light flickering across his features. "Lots of dimensions, especially when they need to find the Hessian. What if… what if we made them deal with *infinitely* many dimensions?"…
>
> It was Banach who finally spoke. "David, they already know about the vector space for the polynomials".
>
> Hilbert smirked. "Who said anything about *countably* infinite?". More silence, then glances, then grins.
>
> It was Riesz's voice which next broke the silence. "And we can make them do analysis in that space. And linear algebra, but not the easy parts. Of course, they'll need to also deal with complex numbers. Sprinkle a little topology and abstract algebra on top, because... they deserve –"
>
> "Frigyes, some of them might actually be able to do that. We need more." After a pause, Fredholm continued: "We'll tell them that they only need to know basic calculus."

A Friendly Approach to Functional Analysis

I didn't actually find the book overly hard (it took me seven days to complete, which is how long it took for my first book, *Naïve Set Theory*), although there were some parts I skipped due to unclear exposition. it's actually one of my favorite books I've read in a while – it's for sure my favorite since the last one. That said, I'm very glad I didn't attempt this early in my book-reading journey.

My brain won't stop line to me

Some part of me *insisted* that the left-shift mapping

$$(x_1, x_2, \ldots) \mapsto (x_2, x_3, \ldots) : \ell^\infty \to \ell^\infty$$

is "non-linear" because it incinerates x_1! But wait, brain, this totally *is* linear, and it's also continuous with respect to the ambient supremum norm!

Formally, a map T is linear when $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$.

Informally, linearity is about being able to split a problem into small parts which can be solved individually. It doesn't have to "look like a line", or something. In fact, lines[1] $y = mx$ are linear *because* putting in Δx more x gets you $m \cdot \Delta x$ more y!

Linearity and continuity

Two things surprised me.

First, a(n infinite-dimensional) linear function can be discontinuous. (?!)

Second, a linear function T is continuous if and only if it is bounded; that is, there is an $M > 0$ such that $\forall x, x_0 : ||T(x - x_0)|| \leq M||x - x_0||$.

- The **if** is easy: this is just Lipschitz continuity, which obviously implies normal continuity.
- The other direction follows because the continuity implies that for $\epsilon := 1$, we can bound how much it's expanding the volume of some δ-ball and then apply linearity.

What the hell are functional derivatives?

Derivatives tell you how quickly a function is changing in each input dimension. In single-variable calculus, the derivative of a function $f : R \to R$ is a function $f' : R \to R$.

In multi-variable calculus, the derivative of a function $g : R^n \to R$ is a function $g' : R^n \to R^n$ – for a given n-dimensional input vector, the real-valued output of g can change differently depending on in which input dimension change occurs.

You can go even further and consider the derivative of $h : R^n \to R^m$, which is the function $h' : R^n \to R^{n \times m}$ – for a given n-dimensional input vector, h again can change its vector-valued output differently depending on in which input dimension change occurs.

But what if we want to differentiate the following function, with domain $C[a, b]$ and range R:

$$L(f) := \int_0^1 (f(t))^2 dt.$$

How do you differentiate with respect to a function? I'm going to claim that

$$L_f'(g) = \int_0^1 2f(t)g(t)dt.$$

It's not clear why this is true, or what it even means. Here's an intuition: at any given point, there are uncountably many partial derivatives in the function space $C[a, b]$ – there are many, many "directions" in which we could "push" a function f around. $L_f'(g)$ gives us the partial derivative at f with respect to g.

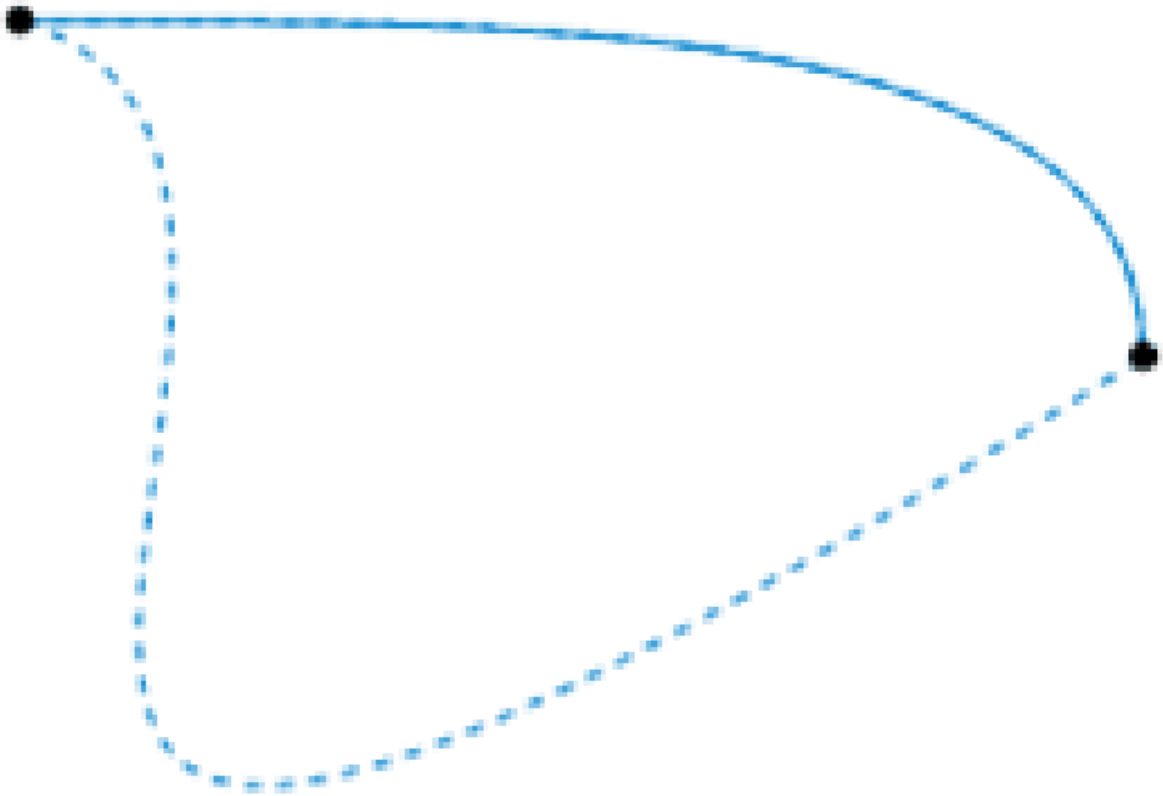
This concept is important because it's what you use to prove e.g. that a line is the shortest continuous path between two points.

Below is an exchange between me (in plain text) and TheMajor (quoted text), reproduced and slightly edited with permission.

I'm having trouble understanding functional derivatives. I'm used to thinking about derivatives as with respect to time, or with respect to variations along the input dimensions. But when I think about a derivative on function space, I'm not sure what the "time" is, even though I can think about the topology and the neighborhoods around a given function.

And I know the answer is that there isn't "time", but I'm not sure what there *is*.

An interesting concept that comes to mind is thinking about a functional derivative with respect to e.g. a straight-line homotopy, where you really *could* say how a function is changing at every point with respect to time. But I don't think that's the same concept.



The concept is as follows:

Let's say we have some (a priori non-linear) map L, which takes a function as an input and gives a number as an output. I.e. it maps from a vector space X of functions to the complex numbers C. Now fix a function $f \in X$, and a second function $g \in X$. We can then consider the 1-dimensional linear subspace $f + Cg := \{f + \lambda g : \lambda \in C\}$. The map L on this subspace is just a normal map, and if it is differentiable at the point f in this subspace then its derivative is called *the functional derivative of* L *at* f *with respect to* g.

By normal map, is that something like a normal operator?

sorry, I didn't mean normal in a technical context. Since the subspace I introduced is one-dimensional (as a complex vector space), and it maps to the complex numbers as well, we have good old introduction to complex analysis derivatives here. If you like you can work with reals instead of complex variables too, in which case it would be the familiar real derivative.

Wouldn't it still output a function, g′ maybe? wait. Would the derivative wrt λ just be g?

> there is no derivative with respect to λ.

ah ya. duh (ETA: my brain was still acting as if differentiation had to be from the real numbers to the real numbers, so it searched for a real/complex number in the problem formalization and found λ.)

> let me know if this part is clear, because unfortunately its the next few steps where it gets really confusing.

Unfortunately, I don't think it's clear yet. So I see how this is a one-dimensional subspace,[2] because it's generated by one basis function (g).

But I don't see how this translates to a normal complex derivative, in particular, I don't quite understand what the range of this function is.

> No problem, and it's very good that you share that it's unclear. The range of L is the complex numbers, L maps from X (our vector space of functions) to C (the complex numbers).

I guess I'm confused why we're using that type signature if we're taking a derivative on the whole function – but maybe that'll be clear after I get the rest.

> that is exactly the heart of the confusion surrounding functional derivatives, and we'll have to get there in a few steps.we'll start with defining functional derivatives for easy maps, i.e. the ones that take on complex values, and then work towards more complicated settings.
>
> so back to the example above; we have a vector space X (our 'function space'), we have a (possibly non-linear) map $L : X \to C$. we will now introduce the derivative of L at f with respect to g, with $f, g \in X$. This derivative is just a complex number.
>
> To find this we consider the 1-dimensional subspace f + Cg that I introduced above, and we note that the map from C to this subspace, given by $\lambda \mapsto f + \lambda g$, is a bijection that goes through f at 0. this gives us a map from C to C, by sending λ to L(f + λg). We take the derivative of that at $\lambda = 0$, and that is *the derivative of* L *at* f *with respect to* g.

Okay, that makes sense so far.

Nice 😀 this map has a few properties that I just want to remark and then ignore. For example it need not be linear in f (which makes sense, since f is only the point we're evaluating at). And by doing some work with chain rules it does have some linear properties in g.

now there are two ways in which we can make this story complicated again, and most authors do both simultaneously.

Firstly we can try to extend the "derivative of L at f wrt g" to something like "derivative of L at f". We'll do this first. Secondly we can try to take a different map, say M, which maps from X into another vector space Y (instead of the complex numbers). We can then try and define a derivative of M at f wrt g.

The first step is conceptually simple, but formally and computationally very difficult. Given a point $f \in X$ and our map L from before, we can simply say that "the derivative of L at f" is the map that sends $g \in X$ to "the derivative of L at f with respect to g". So "the derivative of L at f" is a map from X to C.

this is formally difficult because usually you want this derivative to have some nice properties, but because it was defined pointwise it's very difficult to establish this! Frequently these derivatives are not continuous, and mathematicians resort to horrible tricks (like throwing out a bunch of points of the domain X on which our derivative is annoying) to recover some structure here.

So, given some arbitrary function $L : X \to C$ which is "differentiable" at f, we define a function $L_f' : g \mapsto$ (derivative of L at f with respect to g)?

yes, exactly.

You could even maybe think of each input g as projecting the derivative of L at f? Or specifying one of many possible directions.

Yes, this is 100% correct. This is related to the "nice linear properties in g" that I mentioned above

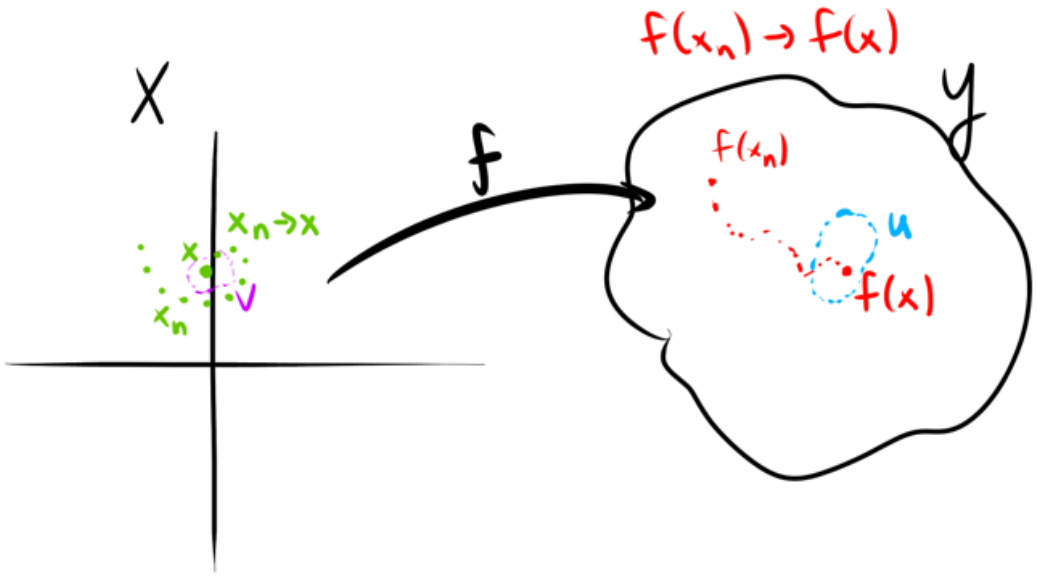
I also stated that this is computationally difficult. This is actually quite funny - the best way to find "The derivative of L at f" is to take a 'test function' $g \in X$ (arbitrarily), compute (the derivative of L at f with respect to g), and then tahdah,

> you have now found the map that sends g to (the derivative of L at f wrt g), i.e. exactly what you were looking for.

this sounds pretty computationally easy? Or are you calculating L' for a general test function g, in which case, how do you get any nontrivial information out of that?

> Yes, you need to calculate it for a general test function.
>
> also something that may help with gaining insight: in multivariable calculus (lets say 2 dimensions, that's already plenty difficult) there is a clear divide between the [existence of a partial derivative of a function at a point] and [the function being differentiable at that point].

ETA: Back in my *Topology* review, I discussed a similar phenomenon: continuity in multiple input dimensions requires not just continuity in each input variable, but in *all* sequences converging to the point in question:



"Continuity in the variables says that paths along the axes converge in the right way. But for continuity overall, we need all paths to converge in the right way. Directional continuity when the domain is R is a special case of this: continuity from below and from above if and only if continuity for all sequences converging topologically to x."

Similarly, for a function to be differentiable, the existence of all of its partial derivatives isn't enough – you need derivatives for every possible approach to the point in question. Here, the existence of all of the partials automatically guarantees the derivatives for every possible approach, because there's a partial for every function.

here we have the same, except we have (in an infinite-dimenional function space X) infinitely many 'partial derivatives'. so from that point of view it's not that surprising that a function "having a derivative at f" is actually quite rare/complicated.

yeah, because L' has to exist for... all g? That seems a little tough.

It exists for all g, and then L_f' exists as a formal map. But usually you want something stronger, for example that $L_f' : X \to C$ is continuous.

as an important but relatively trivial aside: if L is a linear map, then L_f' does not actually depend on f. So usually it is just called "the derivative of L" instead of "the derivative of L at f". This is confusing, because for non-linear L there is also something called "the derivative of L", namely "the map that sends f to [the derivative of L at f]".

hm. That's because of the definition of linearity, right? it's a homomorphism for both the operations of addition and scalar multiplication... Wait, I intuitively understand why linearity means it's the same everywhere, but I'm having trouble coming up with the formal justification...

Yes, the point is that when we look at the definition of "derivative of L at f wrt g" that is given by $\lim_{\lambda \to 0} \frac{L(f+\lambda g)-L(f)}{\lambda}$...

ah, got it!

ok, so this was all the first way to make it confusing again. Ready for the second?

I'm ready to be reconfused.

Ok, so now let's pick a range not inside the complex numbers C, but inside a second normed vector space Y. So we have a map $M : X \to Y$, not necessarily linear. Again fix points $f, g \in X$. We are going to define the derivative of M at f wrt g.

so we repeat our trick from before, consider the map from C via X to Y given by $\lambda \mapsto M(f + \lambda g)$. We wish to differentiate it at $\lambda = 0$.

unfortunately, its image is now in Y, not in C, so we don't really know what the derivative means. But because Y is a normed vector space, the expression $\frac{M(f+\lambda g)-M(f)}{\lambda}$ makes sense for all non-zero λ.

if this function can be continuously extended to $\lambda = 0$ then we define its image at 0 as the derivative of M at f wrt g. Note that this notion of continuity has to do with the norm of Y.

this is now a vector in Y, so if this works we have: [the derivative of M at f wrt g] which is an element of Y, [the derivative of M at f] which is a (linear! usually horrible and not continous!) map from X to Y.

btw if the "continuously extending" part is new, you can also just think of it as the limit of that fraction as λ approaches 0. The only point is that (as long as we're working with complex vector spaces) there are a lot of different ways for λ to approach 0, and it has to work for all of them.

if we're working over the reals its simply the notion of "right limit" and "left limit" (the only two ways to approach 0 in R) that you may have seen before, except that the convergence is now happening in Y.

Other notes

- The operator norm is really cool.
- Linear combinations always involve finitely many terms, but using the orthonormal basis of an infinite dimensional space, you can take the limit as $n \to \infty$.
- I was really happy to see watered-down versions of symmetry/conservation law correspondences (aka Noether's theorem). Can't wait to learn the real version.

Final thoughts

The book is pretty nice overall, with some glaring road bumps – apparently, the Euler-Lagrange equation is one of the most important equations of all time, and Sasane barely spends any effort explaining it to the reader!

And if I didn't have the help of TheMajor, I wouldn't have understood the functional derivative, which, in my opinion, was *the* profoundly important insight I got from this book. My models of function space structure feel qualitatively improved. I can look at a Fourier transform and see what it's doing – I can *feel* it, to an extent. Without a doubt, that single insight makes it all worth it.

Forward

I'm probably going to finish up an epidemiology textbook, before moving on to complex analysis, microeconomics, or... something else – who knows! If you're interested in taking advantage of quarantine to do some reading, feel free to reach out and maybe we can work through something together. 🙂

---

1. Lines $y = mx + b$ ($b \neq 0$) aren't actually linear functions, because they don't go through the origin. Instead, they're affine. ↩

2. To be more specific, $f + Cg := \{f + \lambda g : \lambda \in C\}$ is often an affine subspace, because the zero function is not necessarily a member. ↩

Problem relaxation as a tactic

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

It's easier to make your way to the supermarket than it is to compute the fastest route, which is yet easier than computing the fastest route for someone running backwards and doing two and a half jumping jacks every five seconds and who only follows the route p percent of the time. Sometimes, constraints are necessary. Constraints come with costs. Sometimes, the costs are worth it.

Aspiring researchers trying to think about AI alignment might[1] have a failure mode which goes something like... this:

> Oh man, so we need to solve both outer and inner alignment to build a superintelligent agent which is competitive with unaligned approaches and also doesn't take much longer to train, and also we have to know this ahead of time. Maybe we could use some kind of prediction of what people want... but wait, there's also problems with using human models! How can it help people if it can't model people? Ugh, and what about self-modification?! How is this agent even reasoning about the universe from inside the universe?

The aspiring researcher slumps in frustration, mutters a curse under their breath, and hangs up their hat – "guess this whole alignment thing isn't for me...". And isn't that so? All their brain could do was pattern-match onto already-proposed solutions and cached thinking.

There's more than one thing going wrong here, but I'm just going to focus on one. Given that person's understanding of AI alignment, this problem is *wildly* overconstrained. Whether or not alignment research is right for them, there's just no way that anyone's brain is going to fulfill this insane solution request!

Sometimes, constraints are necessary. I think that the alignment community is pretty good at finding plausibly necessary constraints. Maybe some of the above *aren't* necessary – maybe there's One Clever Trick you come up with which obviates one of these concerns.

Constraints come with costs. Sometimes, the costs are worth it. In this context, I think the costs are very much worth it. Under this implicit framing of the problem, you're pretty hosed if you don't get even outer alignment right.

However, even if the real problem has crazy constraints, that doesn't mean you should immediately tackle the fully constrained problem. I think you should often relax the problem first: eliminate or weaken constraints until you reach a problem which is still a little confusing, but which you can get some traction on.

> Even if you know an unbounded solution to chess, you might still be 47 years away from a bounded solution. But if you can't state a program that solves the problem in principle, you are in some sense confused about the nature of the cognitive work needed to solve the problem. If you can't even solve a problem given infinite computing power, you definitely can't solve it using bounded computing power. (Imagine Poe trying to write a chess-playing program before he'd had the insight about search trees.)

~ *The methodology of unbounded analysis*

Historically, I tend to be too slow to relax research problems. On the flipside, *all of my favorite research ideas were directly enabled by problem relaxation*. Instead of just telling you what to do and then having you forget this advice in five minutes, I'm going to paint it into your mind using two stories.

Attainable Utility Preservation

It's spring of 2018, and I've written myself into a corner. My work with CHAI for that summer was supposed to be on impact measurement, but I *inconveniently* posted a convincing-to-me argument that impact measurement cannot admit a clean solution:

> I want to penalize the AI for having side effects on the world.[2] Suppose I have a function which looks at the consequences of the agent's actions and magically returns all of the side effects. Even if you have this function, you still have to assign blame for each effect – either the vase breaking was the AI's fault, or it wasn't.
>
> If the AI penalizes itself for everything, it'll try to stop people from breaking vases – it'll be clingy. But if you magically have a model of how people are acting in the world, and the AI magically only penalizes itself for things which are its fault, then the AI is incentivized to blackmail people to break vases in ways which don't *technically* count as its fault. Oops.

Summer dawned, and I occupied myself with reading – lots and lots of reading. Eventually, enough was enough – I wanted to figure this out. I strode through my school's library, markers in my hand and determination in my heart. I was determined not to leave before understanding *a)* exactly why impact measurement is impossible to solve cleanly, or *b)* how to solve it.

I reached the whiteboard, and then – with adrenaline pumping through my veins – I realized that I had *no idea* what this "impact" thing even is. Oops.

I'm staring at the whiteboard.

A minute passes.

59 more minutes pass.

I'd been thinking about how, in hindsight, it was so important that Shannon had first written a perfect chess-playing algorithm which required infinite compute, that Hutter had written an AGI algorithm which required infinite compute. I didn't know how to solve impact under all the constraints, but what if I assumed something here?

> What if I had infinite computing power? No... Still confused, don't see how to do it. Oh yeah, and what if the AI had a perfect world model. Hm... *What if we could write down a fully specified utility function which represented human preferences? Could I measure impact if I knew that?*

The answer was almost trivially obvious. My first thought was that negative impact would be a decrease in true utility, but that wasn't quite right. I realized that impact measure needs to also capture decrease in ability to achieve utility. That's an optimal value function... So the negative impact would be the decrease in attainable utility for human values![3]

> Okay, but we don't and won't know the "true" utility function. What if... we just penalized shift in all attainable utilities?

I then wrote down The Attainable Utility Preservation Equation, more or less. Although it took me a few weeks to believe and realize, that equation solved all of the impact measurement problems which had seemed so insurmountable to me just minutes before.[4]

Formalizing Instrumental Convergence

It's spring of 2019, and I've written myself into a corner. My first post on AUP was confusing – I'd failed to truly communicate what I was trying to say. Inspired by *Embedded Agency*, I was planning an illustrated sequence of my own.

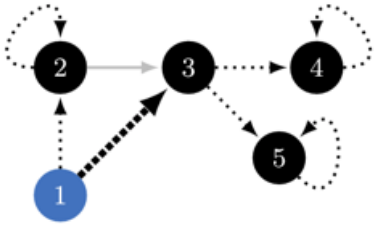
I was working through a bit of reasoning on how your ability to achieve one goal interacts with your ability to achieve seemingly unrelated goals. Spending a lot of money on red dice helps you for the collecting-dice goal, but makes it harder to become the best juggler in the world. That's a weird fact, but it's an *important* fact which underlies much of AUP's empirical success. I didn't understand why this fact was true.

At an impromptu presentation in 2018, I'd remarked that "AUP wields instrumental convergence as a weapon against the alignment problem itself". I tried thinking about it using the formalisms of reinforcement learning. Suddenly, I asked myself

> Why is instrumental convergence even a thing?

I paused. I went outside for a walk, and I paced. The walk lengthened, and I still didn't understand why. Maybe it was just a "brute fact", an "emergent" phenomenon – nope, not buying that. There's an explanation somewhere.

I went back to the drawing board – to the whiteboard, in fact. I stopped trying to understand the general case and I focused on specific toy environments. I'm looking at an environment like this



and I'm thinking, most agents go from 1 to 3. "Why does my brain think this?", I asked myself. Unhelpfully, my brain decided not to respond.

I'm staring at the whiteboard.

A minute passes.

29 more minutes pass.

I'm reminded of a paper my advisor had me read for my qualifying exam. The paper talked about a dual formulation for reinforcement learning environments, where you consider the available trajectories through the future instead of the available policies. I take a picture of the whiteboard and head back to my office.

I run into a friend. We start talking about work. I say, "I'm about 80% sure I have the insight I need – this is how I felt in the past in situations like this, and I turned out to be right".

I turned out to be right. I started building up an entire theory of this dual formalism. Instead of asking myself about the general case of instrumental convergence in arbitrary computable environments, I considered small deterministic Markov decision processes. I started proving everything I could, building up my understanding piece by piece. This turned out to make all difference.

Half a year later, I'd built up enough theory that I was able to explain a great deal (but not everything) about instrumental convergence.

Conclusion

Problem relaxation isn't always the right tactic. For example, if the problem isn't well-posed, it won't work well – imagine trying to "relax" the "problem" of free will! However, I think it's often the right move.

The move itself is simple: consider the simplest instance of the problem which is still confusing. Then, make a ton simplifying assumptions while still keeping part of the difficulty present – don't assume away all of the difficulty. Finally, tackle the relaxed problem.

In general, this seems like a skill that successful researchers and mathematicians learn to use. MIRI does a lot of this, for example. If you're new to the research game, this might be one of the crucial things to pick up on. Even though I detailed how this has worked for me, I think I could benefit from relaxing more.

The world is going to hell. You might be working on a hard (or even an impossible) problem. We plausibly stand on the precipice of extinction and utter annihilation.

Just relax.

This is meant as a reference post. I'm not the first to talk using problem relaxation in this way. For example, see The methodology of unbounded analysis.

---

1. This failure mode is just my best guess – I haven't actually surveyed aspiring researchers. ↩

2. The "convincing-to-me argument" contains a lot of confused reasoning about impact measurement, of course. For one, thinking about side effects is *not* a good way of conceptualizing the impact measurement problem. ↩

3. The initial thought wasn't as clear as "penalize decrease in attainable utility for human values" – I was initially quite confused by the AUP equation. "What the heck *is* this equation, and how do I break it?".

It took me a few weeks to get a handle for why it seemed to work so well. It wasn't for a month or two that I began to understand what was actually going on, eventually leading to the *Reframing Impact* sequence. However, for the reader's convenience, I whitewashed my reasoning here a bit. ↩

4. At first, I wasn't very excited about AUP – I was new to alignment, and it took a lot of evidence to overcome the prior improbability of my having actually found something to be excited about. It took several weeks before I stopped thinking it likely that my idea was probably secretly and horribly bad.

   However, I kept staring at the strange equation – I kept trying to break it, to find some obvious loophole which would send me back to the drawing board. I never found it. Looking back over a year later, AUP does presently have loopholes, but they're not obvious, nor should they have sent me back to the drawing board.

   I started to get excited about the idea. Two weeks later, my workday was wrapping up and I left the library.

   > Okay, I think there's about a good chance that this ends up solving impact. If I'm right, I'll want to have a photo to commemorate it.

   I turned heel, descending back into the library's basement. I took the photograph. I'm glad that I did.

   Discovering AUP was one of the happiest moments of my life. It gave me confidence that I could think, and it gave me some confidence that we can *win* – that we can solve alignment. ↩

Insights from Euclid's 'Elements'

Presumably, I was taught geometry as a child. I do not remember.

Recently, I'd made my way halfway through a complex analysis textbook, only to find another which seemed more suitable and engaging. Unfortunately, the exposition was geometric. I knew something was wrong – I knew something had to change – when, asked to prove the similarity of two triangles, I got stuck on *page 7*.

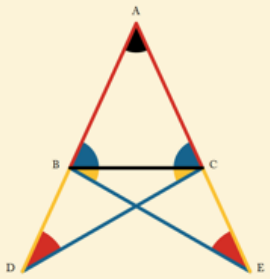
I'd been reluctant to tackle geometry, and when authors reasoned geometrically, I'd find another way to understand. Can you blame me, when most geometric proofs look like *this*?

LET the equal ſides AB and AC be produced through the extremities BC, of the third ſide, and in the produced part BD of either, let any point D be aſſumed, and from the other let AE be cut off equal to AD (B. 1. pr. 3). Let the points E and D, ſo taken in the produced ſides, be connected by ſtraight lines DC and BE with the alternate extremities of the third ſide of the triangle.

In the triangles DAC and EAB the ſides DA and AC are reſpectively equal to EA and AB, and the included angle A is common to both triangles. Hence (B. 1. pr. 4.) the line DC is equal to BE, the angle ADC to the angle AEB, and the angle ACD to the angle ABE; if from the equal lines AD and AE the equal ſides AB and AC be taken, the remainders BD and CE will be equal. Hence in the triangles BDC and CEB, the ſides BD and DC are reſpectively equal to CE and EB, and the angles D and E included by thoſe ſides are alſo equal. Hence (B. 1. pr. 4.) the angles DBC and ECB, which are thoſe included by the third ſide BC and the productions of the equal ſides AB and AC are equal. Alſo the angles DCB and EBC are equal if thoſe equals be taken from the angles DCA and EBA before proved equal, the remainders, which are the angles ABC and ACB oppoſite to the equal ſides, will be equal.

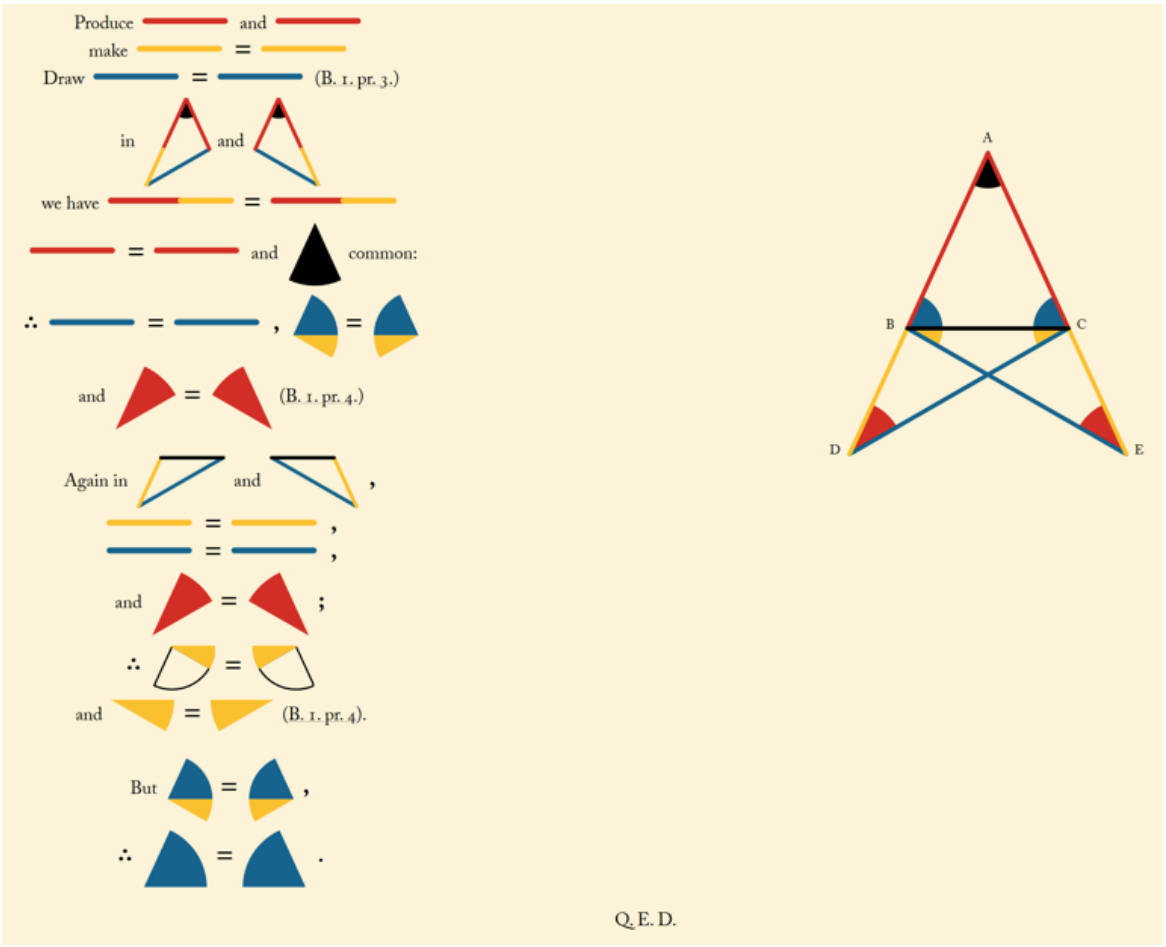
Therefore in an iſoſceles triangle, &c.

Q. E. D.



Distasteful. In a graph with n vertices, you'd need to commit $O(n^3)$ things to memory (e.g. triangles, angles) in order to read the proof without continually glancing at the illustration. In a normal equation with n variables, it's O(n).

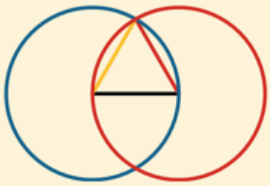
Sometimes, we just need a little beauty to fall in love.



Welcome to Oliver Byrne's rendition of Euclid's *Elements*, digitized and freely available online.

PROPOSITION I. PROBLEM.

ON *a given finite ſtraight line* (——) *to deſcribe an equilateral triangle.*



Deſcribe ◯ and ◯ (poſtulate 3.); draw —— and —— (poſt. 1.). then △ will be equilateral.

For —— = —— (def. 15.);
and —— = —— (def. 15.),
∴ —— = —— (axiom. 1.);

and therefore △ is the equilateral triangle required.

Q. E. D.

Elements

> Propoſitions are placed before a ſtudent, who though having a ſufficient underſtanding, is told juſt as much about them on entering at the very threſhold of the ſcience, as gives him a prepoſſeſſion moſt unfavourable to his future ſtudy of this delightful ſubject; or “the formalities and paraphernalia of rigour are ſo oſtentatiouſly put forward, as almoſt to hide the reality. Endleſs and perplexing repetitions, which do not confer greater exactitude on the reaſoning, render the demonſtrations involved and obſcure, and conceal from the view of the ſtudent the conſecution of evidence.”
>
> Thus an averſion is created in the mind of the pupil, and a ſubject ſo calculated to improve the reaſoning powers, and give the habit of cloſe thinking, is degraded by a dry and rigid courſe of inſtruction into an unintereſting exerciſe of the memory.
>
> ~ Oliver Byrne

Equality and Similarity

Old mathematical writing lacks modern precision. Euclid says that two triangles are "equal", without specifying what that means. It means that one triangle can be turned into another via an isometric transformation. That is, if you rotate, translate, and/or reflect triangle A, it turns into triangle B.

Similarity is a bit more lenient, because you can rescale as well:

My favorite characterization of similarities is:

As a map $f : R^n \to R^n$, a similarity of ratio r takes the form $f(x) := rAx + t$, where $A \in O_n(R)$ is an $n \times n$ orthogonal matrix and $t \in R^n$ is a translation vector.

The only difference compared to congruence is that congruence requires $r = 1$.

Synthetic/analytic

I find it strange that Euclid got so far by axiomatizing informal notions without any grounding in formal set theory (e.g. ZFC). I mean, you'd get *absolutely blown away* if you tried to pull these shenanigans in topology. But apparently, Euclidean geometry is sufficiently well-behaved that it basically matches our intuitions without much qualification?

Area invariance

Book 1, proposition 35:

PROPOSITION XXXV. THEOREM.

PARALLELOGRAMS *on the same base, and between the same parallels, are (in area) equal.*

This says: suppose you draw two parallel lines, and then make a dash of length 2 on each line. Then, make another dash of length 2 on the upper line. The two parallelograms so defined have equal area. This is clarified in the next theorem.

PROPOSITION XXXVI. THEOREM.

PARALLELOGRAMS ( *and* ) *on equal bases, and between the same parallels, are equal.*

If you take one of the dashes and slide it around on the upper parallel line, the resultant parallelograms all have the same area. I thought this was cool.

Notes

There aren't any exercises; instead, I tried to first prove the theorems myself.

Book III treats circles, with wonderful results on arcs and their relation to angles. I search for a snappy example, a gem of an insight to share, but my words fail me. It's just good.

I read books I, III, IV, and skimmed II. Not all books of the Elements are about plane geometry; some are archaic introductions to number theory, for example. Those looking to learn number theory would do much better with the gorgeous *Illustrated Theory of Numbers*.

Forward

Elements is a *tour de force*. Theorem, theorem, problem, theorem, all laid out in confident succession. It was not always known that from simple rules you could rigorously deduce beautiful facts. It was *not always known* that you could start with so little, and end with so much.

Before I found this resource, I'd checked out several geometry books, all of which seemed bad. To salt the wound, many books were explicitly aimed at middle-schoolers. This... was a bit of a blow.

However, it doesn't matter when something is normally presented. If you don't know something, you don't know it, and there's nothing wrong with learning it. Even if you feel late. Even if you feel sheepish.

Against completionism

I'm glad I didn't read all of the books, even though they're beautiful. I'd picked up a bad "completionist" habit – if I don't read the whole book, obviously I haven't completed it, and obviously I'm not allowed to make a post about it. Of course.

But I'm trying to pick up useful skills, to expand the types of qualitative reasoning available to me, to get the most benefit per unit of reading. I stopped because I have what I need for my complex analysis book.

Read around

Reading relevant Wikipedia pages / other textbooks helps me cross-examine my knowledge. It also helps connect the new knowledge to existing knowledge. For example, I now have a wonderfully enriched understanding of the geometric mean.

Over time, as you expand and read more books, you'll find yourself reading faster and faster, understanding more and more subsections. I don't recommend learning new areas via Wikipedia, but it's good reinforcement.

Re-deriving dependencies as a habit

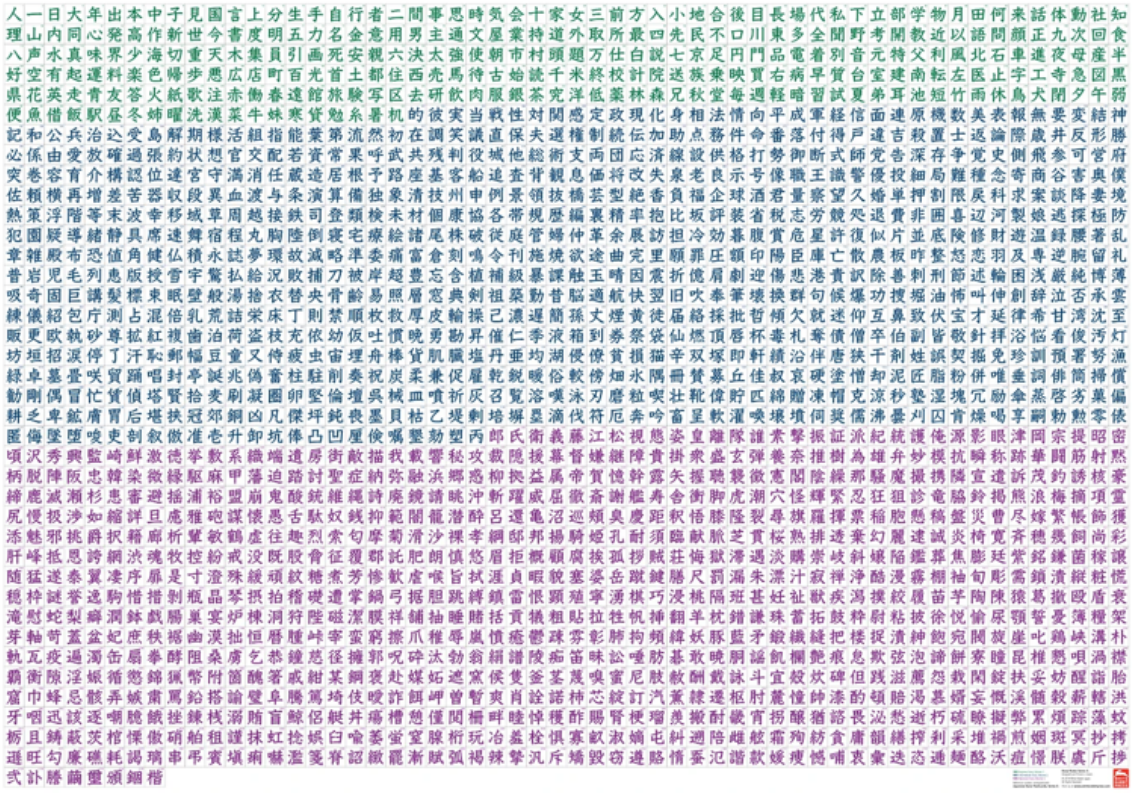
Ever since I learned real analysis, I reflexively reprove all new elementary mathematics whenever I use it. For real analysis, that meant *continually reproving* e.g. $1 + 2 = 2 + 1$ whenever I used that property in a proof. Did it feel silly and tedious? A bit, yes.

But with (this) tedium comes power. I can now regenerate a formal foundation for the real numbers from the Peano axioms, proving the necessary properties about the natural numbers, then the integers, then the rationals, and then the reals, and then complex numbers too. (But please, no quaternions!)

With this habit, you continually ask yourself, "how do I know this?". I think this is a useful subskill of Actually Thinking.

Commemoration

In college, I taught myself a bit of Japanese. Through a combination of spaced repetition software and memory palaces, and over the course of three months, I learned to read the 2,136 standard use characters. After those three months, I proudly displayed this poster on my wall:



I look forward to another beautiful poster.

As the ſenſes of ſight and hearing can be ſo forcibly and inſtantaneously addreſſed alike with one thouſand as with one, *the million* might be taught geometry and other branches of mathematics with great eaſe, this would advance the purpoſe of education more than any thing that *might* be named, for it would teach the people how to think, and not what to think; it is in this particular the great error of education originates.

Lessons I've Learned from Self-Teaching

In 2018, I was a bright-eyed grad student who was freaking out about AI alignment. I guess I'm still a bright-eyed grad student freaking out about AI alignment, but that's beside the point.

I wanted to help, and so I started levelling up. While I'd read Nate Soares's self-teaching posts, there were a few key lessons I'd either failed to internalize or failed to consider at all. I think that implementing these might have doubled the benefit I drew from my studies.

I can't usefully write a letter to my past self, so let me write a letter to you instead, keeping in mind that good advice for past-me may not be good advice for you.

Make Sure You Remember The Content

TL;DR: use a spaced repetition system like Anki . Put in cards for key concepts and practice using the concepts. Review the cards every day without fail. This is the most important piece of advice.

The first few months of 2018 were a dream: I was learning math, having fun, and remaking myself. I read and reviewed about one textbook a month. I was learning how to math, how to write proofs and read equations fluently and think rigorously.

I had so much fun that I hurt my wrists typing up my thoughts on impact measures. This turned a lot of my life upside-down. My wrists wouldn't fully heal for two years, and a lot happened during that time. After I hurt my wrists, I became somewhat depressed, posted less frequently, and read fewer books.

When I looked back in 2019/2020 and asked "when and why did my love for textbooks sputter out?", the obvious answer was "when I hurt my hands and lost my sense of autonomy and became depressed, perchance? And maybe I just became averse to reading that way?"

The obvious answer was wrong, but its obvious-ness stopped me from finding the truth until late last year. It *felt* right, but my introspection had failed me.

The real answer is: when I started learning math, I gained a lot of implicit knowledge, like how to write proofs and read math (relatively) quickly. However, I'm no Hermione Granger: left unaided, I'm bad at remembering explicit facts / theorem statements / etc.

I gained implicit knowledge *but I didn't remember the actual definitions*, unless I actually used them regularly (*e.g.* as I did for real analysis, which I remained quite fluent in and which I regularly use in my research). Furthermore, I think I *coincidentally* hit steeply diminishing returns on the implicit knowledge around when I injured myself.

So basically I'm reading these math textbooks, doing the problems, getting a bit better at writing proofs but not really durably remembering 95% of the content. Maybe part of my subconscious noticed that I seem to be wasting time, that when I come back four months after reading a third of a graph theory textbook, I barely remember the new content I had "learned." I thought I was doing things right. I was doing dozens of exercises and thinking deeply about why each definition was the way it was, thinking about how I could apply these theorems to better reason about my own life and my own research, etc.

I explicitly noticed this problem in late 2020 and thought,

> is there any way I know of to better retain content?

> ... gee, what about that thing I did in college that let me learn how to read 2,136 standard-use Japanese characters in 90 days? you know, Anki spaced repetition, that thing I *never tried* for math because once I tried and failed to memorize dozens of lines of MergeSort pseudocode with it?
>
> hm...

This was the moment I started feeling extremely silly (the exact thought was "there's no possible way that my hand is big enough for how facepalm this moment is", IIRC), but also extremely excited. *I could fix my problem!*

And a problem this was. In early 2020, I had an interview where I was asked to compute

$\int x \log x \, dx$. I was stumped, even though this was simple high school calculus (just integrate by parts!). I failed the interview and then went back to learning algebraic topology and functional analysis and representation theory. You know, nothing difficult like high school calculus.

I was pretty frustrated with myself.

It's not that I didn't understand it. I just didn't remember it, especially on the spot. The worst part was that I had brushed up on calculus *the previous spring*, and I still didn't remember it. Turns out that my brain won't remember material it doesn't use for months on end, even if forgetting that material would be embarrassing.

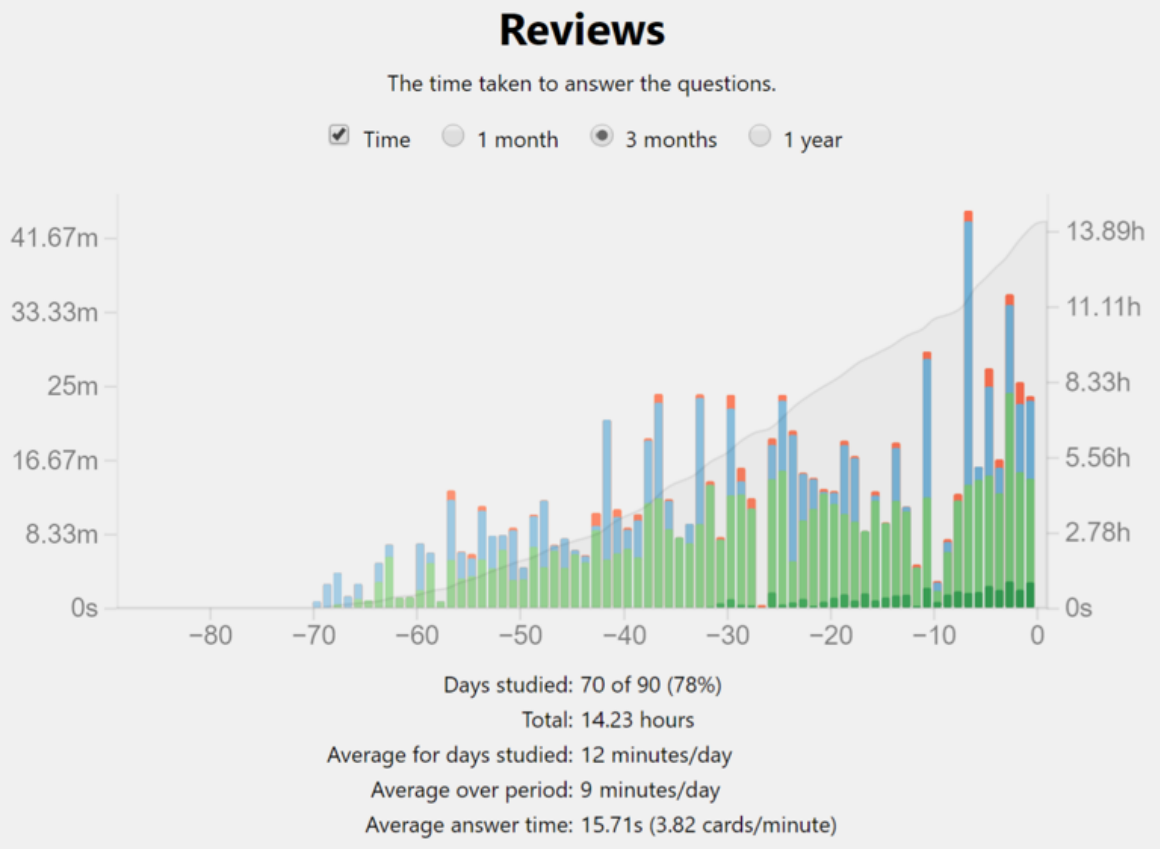
Enter Anki, an amazing spaced repetition system ($20 for iOS, free for computer). The way I like to explain Anki is:

> Anki is a flashcard application into which you can enter a constant number of cards each day while retaining a constant average daily workload. You can add cards each day, without having to study longer and longer to get through all of the cards.

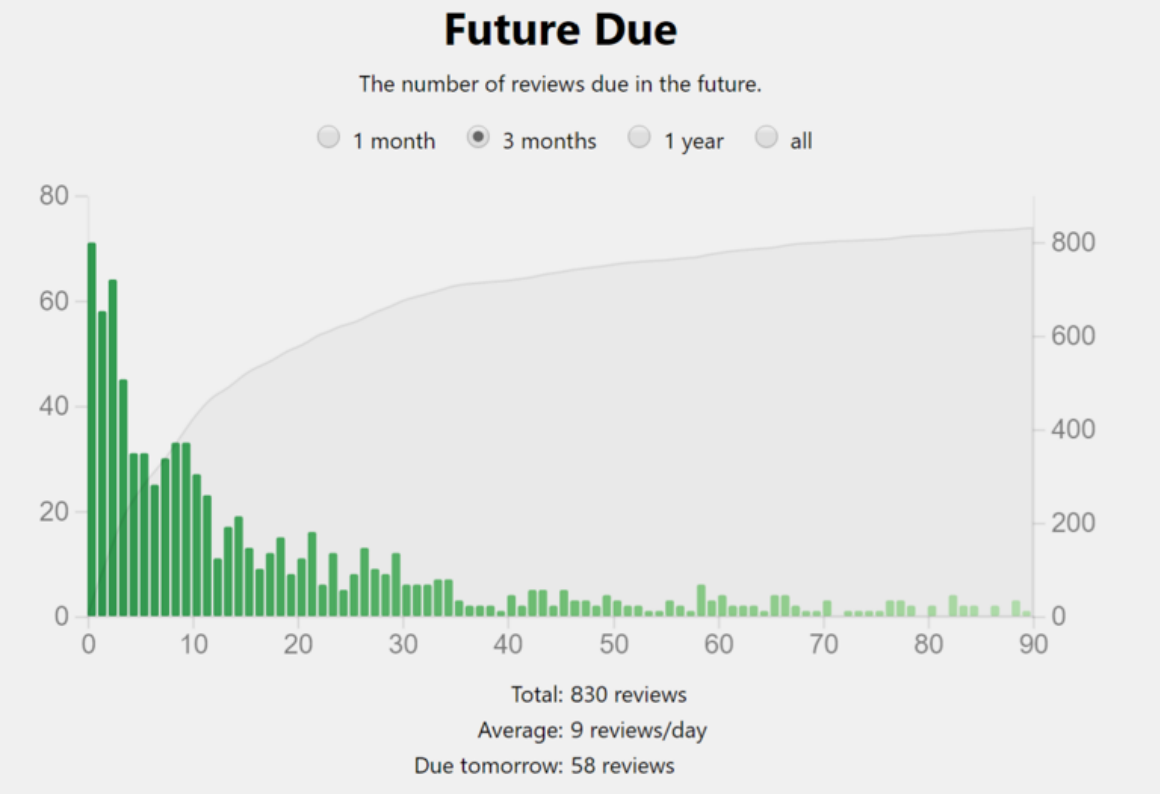
I currently think that unless you have really good memory or you're not learning content you want to remember months from now, you're making a mistake by not using a spaced repetition system. Read Gwern for more on spaced repetition.

Spaced repetition seems especially useful for students. In college, I ran an experiment: for an upper-level French class, I put things I didn't know how to say into Anki, reviewed daily, and otherwise didn't study at all. I got an A.

How powerful might a bright 6th grader grow, were they to use Anki every day for their whole life? The best time to plant a tree may have been in sixth grade, and the second-best time may have been in seventh grade, but you should still plant the tree now rather than never.



I've been using Anki for math for the last 71 days, and I currently have a deck of about 900 cards which I study for ~30 minutes daily. In 2018, I spent about 10 minutes daily reviewing a deck of nearly ten thousand French cards.



If I were to add no more cards, daily reviews drop off quickly.

(Once completed, the reviews in the next few days *will* be pushed into some of the future days, so this projection is slightly optimistic, but you get the point.)

I love Anki, and I was foolish to circumscribe it to language-learning. I now use Anki to remember key concepts from academic talks, LessWrong blogposts, and yes - textbooks. Which I now love again, and which I read for ~an hour daily again, because I'm *actually retaining the content*.

Measure theory, ring theory, random stuff about taxonomy and epidemiology, quantum mechanics, basic physics, deep RL papers, they all go into Anki, and Anki cards go into my brain - and stay there!

What a wonderful time to be alive.

Random Anki tips

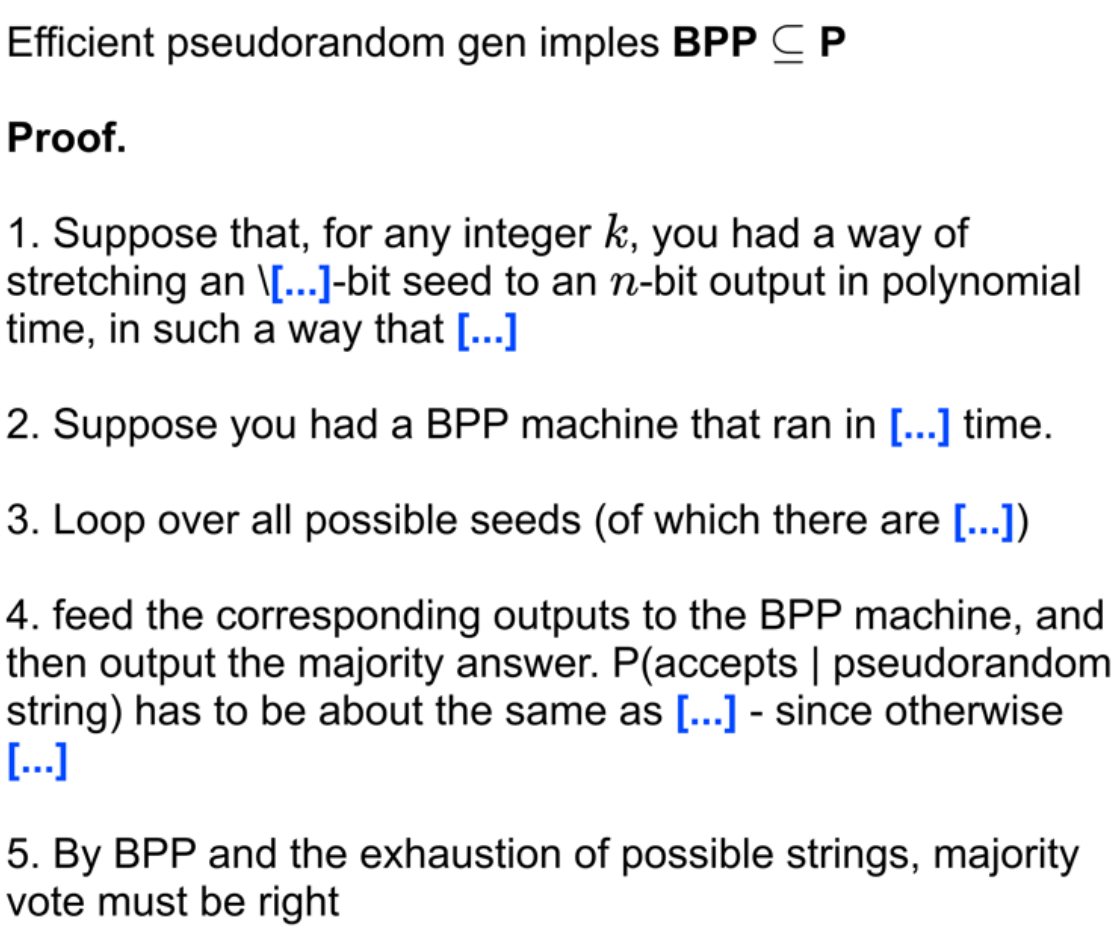
I know quite a bit about how to best use Anki, so if you try this and it doesn't seem to work, please message me instead of banging your head against the wall or giving up!

- **Use cloze deletions on short cards**. Cloze deletions hide a small part of the text, and make you remember it given the rest of the content. They are fast to make and fast to review.
    - The vast, vast majority of my cards are cloze now, and they weren't when I first wrote this post. I think that was a mistake.
- Study every day, preferably at the same time so you aren't always scrambling to get it done before bedtime.
- Sync with AnkiWeb so you don't lose all your cards if your device dies.
- Save time by just screenshotting theorem statements and/or proofs.

    - EDIT: micpie recommends the Mathpix OCR software, which clips text into MathJax code. This works really well, in my experience.
    - Image occlusion is a great add-on.
- On iPad, I like using MarginNote to read. I can just draw rectangles around key parts of the pdf, make cloze deletions in the app, and then export the cards to an Anki deck.

Memorizing definitions can be useful: when reading a text, it saves you from having to constantly check what the concept is. Make sure to include examples to work through - don't just toss in random definitions you're barely interested in and will never think about again.

- Don't just memorize proofs, focus on the key ideas. Don't just memorize definitions, throw in several example problems which are small enough to actually do in your head (or with a scrap of paper).
    - For example, if I'm trying to really ingrain the concept of an efficient pseudorandom number generator, I have cards where I reason about it by completing short proofs:

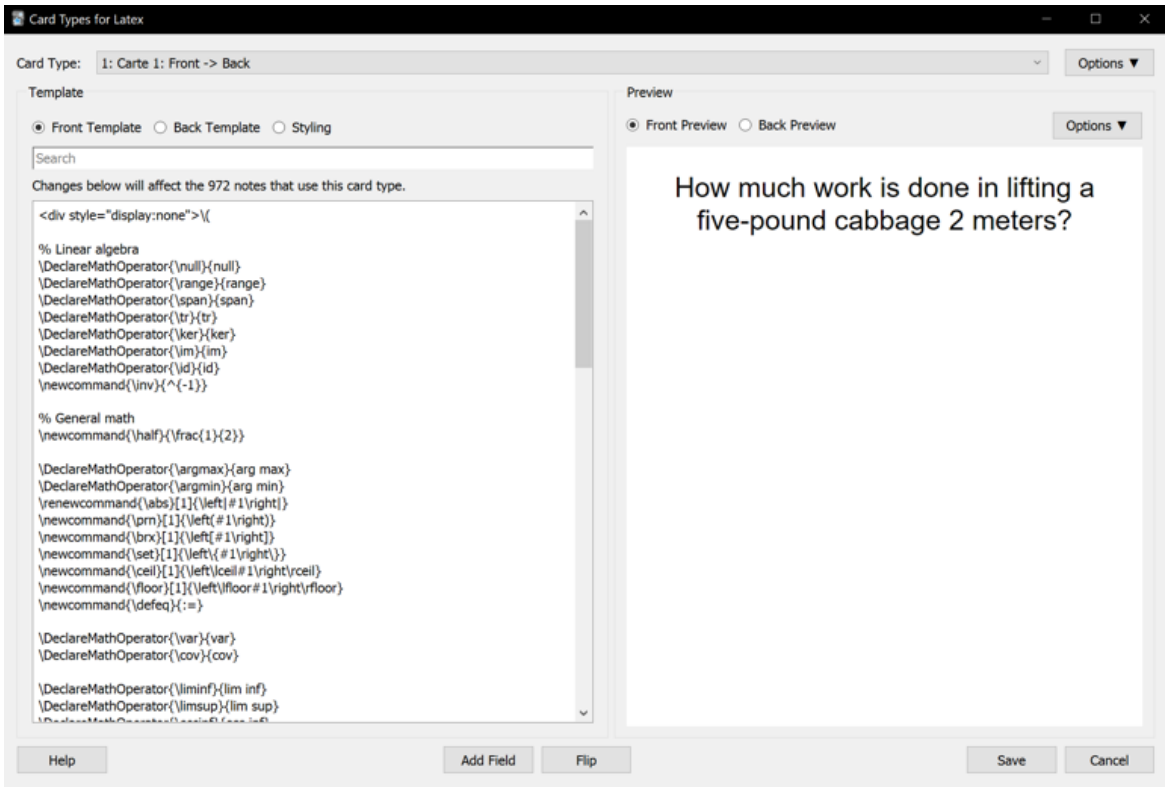


The math here isn't important. The key thing is, I want to remember the "pseudorandom number generator" concept, and so I find an interesting

result and make myself prove it. The proof isn't too long, which is key—just a few cloze deletions. Don't try to memorize essays with Anki.

Am I ever going to actually use this math for my research? Probably not. Doesn't matter. Anki makes it cheap to learn and retain things.

- If you get a card wrong more than 4 times in the first week, it's a bad card. Remake it.
- Use MathJax instead of Latex, because MathJax renders instantly.



If you want custom commands, use the card type editor.

- My trigger for "I should add a new card" is reading something and thinking, "this is a cool concept!"
    - I recommend adding cards liberally. Don't worry about getting the formatting or phrasing perfect at first. Just add cards and you'll develop a taste for what should be added, and how.

Read Several Textbooks Concurrently

TL;DR study several topics at once so that your brain has time to cement the concepts you're learning, before the text builds on those concepts further.

AllAmericanBreakfast's recent post is great, so I'll refer you to that to make this point:

> Wait, what? You want me to make life easier on myself by, instead of studying calculus...studying calculus, linear algebra, *and* statistics all at once?
>
> ~ AllAmericanBreakfast, *The Multi-Tower Study Strategy*

The basic idea is that your brain needs time to really cement a new idea in, and so you should study several topics at once in dependency-heavy areas like mathematics. This

advice matches up with both my recent personal experience and with advice I received early on from Qiaochu Yuan, but which I had unfortunately ignored.

For example, right now I'm reading Nielsen and Chuang's *Quantum Computation and Quantum Information*, Evan Chen's *The Infinite Napkin*, and a ridiculously easy physics book, Kuhn and Noschese's *A Self-Teaching Guide: Basic Physics* (more on this later). Previously, I'd been going back through Wasserman's *All of Statistics* and Pearl's *Causality* after reading Pearl's *Book of Why*.

I always feel like I'm learning something new instead of banging my head against the wall. Sometimes you should just read one book, but if you don't need to cram, I recommend diversifying.

Completing The Whole Textbook Is Usually a Big Waste of Time, Please Don't Do It

TL;DR extract the most useful / central concepts and remember them forever via Anki. This doesn't require grasping every arcanum, every detail of a textbook.

When I started reading textbooks, I completed the whole book. I didn't want to miss a crucial concept. But the thing about crucial concepts is that they pop up everywhere. If you missed a crucial concept, you'll know. You're not going to wake up 20 years from now and be like, "OH NO! I forgot to learn about 'force' when I self-studied physics! And I forgot to learn about injectivity in linear algebra!"

As you read more, you'll get a taste for what's probably important, and what's details you can reference later if need be. You can also ask experts if you should study part of a book - this is one key benefit of having an actual teacher.

But don't complete the whole book and all of its exercises, if you just want to become a polymath. Leverage the Pareto principle, get 80% of the benefit out of the key 20/30/40% of the concepts and exercises, and then move on.

Another reason I used to do this a lot was that I wanted to look good by being able to mention on *e.g.* my resume that I had read the whole book. It's a lot more impressive to say "I read these 10 textbooks" than "I read large parts of this book, and some of this book, and a little of this one, and a lot of this other book." I've found that learning well requires keeping an eye out for these instincts. My brain is not my friend in this battle.

I now often ask myself "am I doing this, at least in part, in order to look good?". Sometimes I answer 'yes', and sometimes I do it anyways - wanting to look good isn't always bad. But there are sometimes things more important than looking good.

Read Easier Textbooks Instead of Struggling Valiantly

TL;DR even though slogging through tough textbooks makes you feel sophisticated and smart, don't.

Imagine I'm learning how to program, but I've never used a computer before. Learning to program *while* learning to operate a computer, will probably take longer than learning to operate a computer and then learning to program. Learning time is superadditive in terms of your ignorance / the dependencies you're missing.

But it's worse than that. Imagine I'm learning quantum mechanics, but I don't know any linear algebra either. I'm now trying to do three things:

1. Learn linear algebra,
2. Learn the formal postulates of quantum mechanics, and
3. Tie all of this into the real world.

Similarly, if I'm trying to learn fluid mechanics without knowing how to manipulate partial differential equations (PDEs), it might look trying to simultaneously

1. Learn PDEs,
2. Learn the physical equations, such as Navier-Stokes, and
3. Tie all of this into the real world to explain what I already know about *e.g.* water and rivers and blood pressure.

But what if instead I picked up some dumb book that doesn't even have any calculus, and let it give me approximate explanations via *e.g.* Archimedes' principle:

> Archimedes' principle states that the upward buoyant force that is exerted on a body immersed in a fluid, whether fully or partially, is equal to the weight of the fluid that the body displaces. (Wikipedia)

I breeze through this book no problem, and I can see how to tie in these laws to explain my intuitive models: "logs float more easily than rocks because rock is denser than wood, and so the buoyant force from Archimedes' principle is enough to support the weight of a log." So I'm taking care of point #3, "tie this content into the real world."

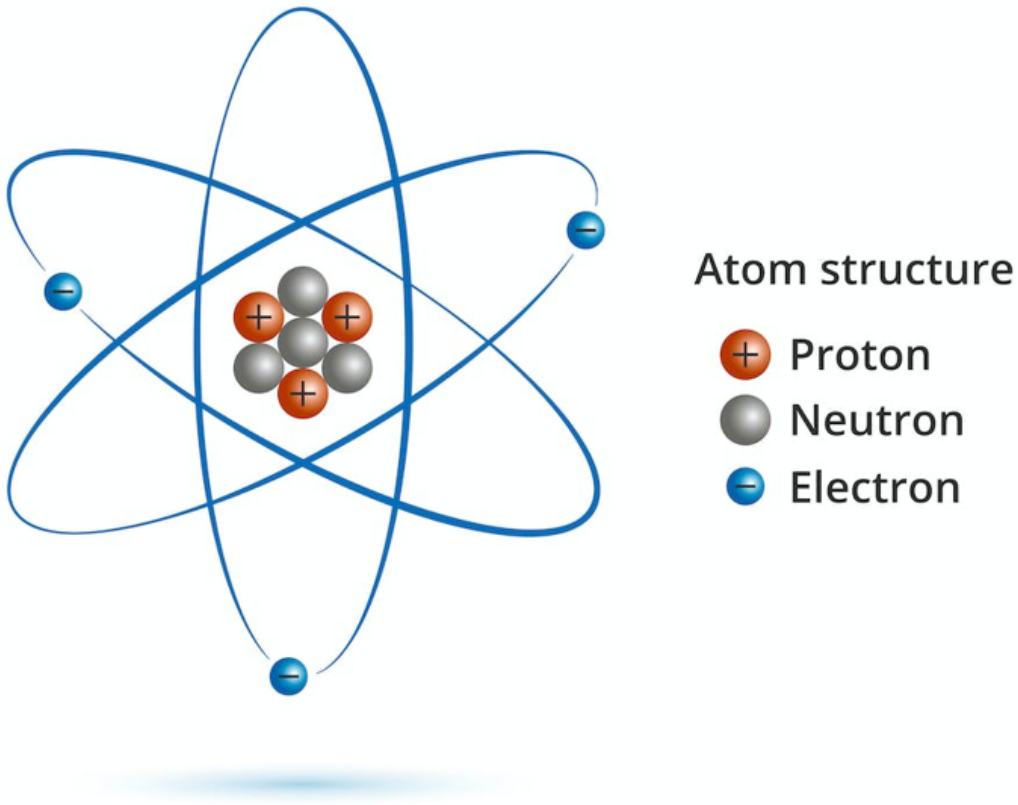
Then, suppose I learn about PDEs and become comfortable with them. Now all I need to do to learn a piece of fluid mechanics is to learn the relevant physical equation, and then think about how it implies things like Archimedes' principle. Crucially, via this method, *I'm only confused about one thing at a time*. Build models in the right order!

Be Comfortable with Approximate Models

TL;DR allow yourself to learn things in order of comprehensibility: don't try to learn general relativity before Newton's law of gravitation.

In 2019, I wanted to learn a bit of chemistry. I got my hands on a high school chemistry textbook. I got stuck on chapter two because I was being too strict about forming gears-level models.

I was stuck because I was thinking about electron shells. The book acknowledged the Bohr model of the atom was wrong, electrons aren't *really* discrete particles orbiting the nucleus:



A wrong model.

But I got nerd-sniped into trying to understand how the electron standing wave only has solutions for certain energy levels, which is a result of quantum mechanics (or so I remember reading). I couldn't understand why, and so the knowledge felt "fake."

It wasn't like I explicitly reasoned "this model is wrong and so I'm not going to keep reading this book", it felt more like... I just wasn't that hungry to learn this, because it wasn't "real." And I gradually stopped returning to that book.

Learn things quickly, note your confusions, and correct them later when the Anki cards show up again. Let yourself learn approximate models of reality.

Conclusion

I knew about the "read easier textbooks" advice already, but I didn't apply it. Perhaps I just didn't recognize a chance to apply it. The same forces of chaos and entropy and madness which prevented my applying *e.g.* Luke Muelhauser's advice, may prevent you from applying this post's advice. If you think any of this advice might help you, I recommend setting up a plan *now* for how and when you'll implement it.

Insights from Modern Principles of Economics

How good are our economists? Look around. On a 20-minute walk to my Berkeley office, I walked past people reeking of urine, past people lying in a dirty sleeping bag on a thin cardboard pad, past some garbage around a tent which housed a child who grew up into an impoverished adult.

Imagine living here.

In what world is this broader system a success story for economics?

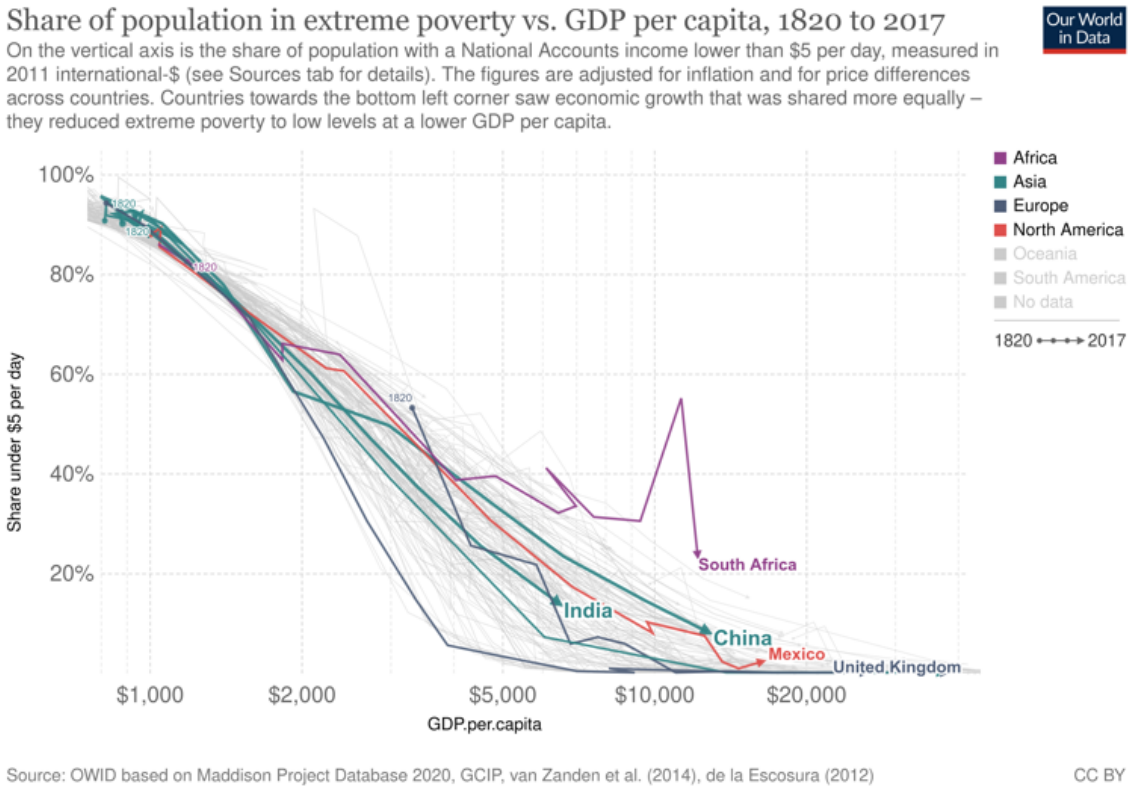
In **this world**.

Economics is important.

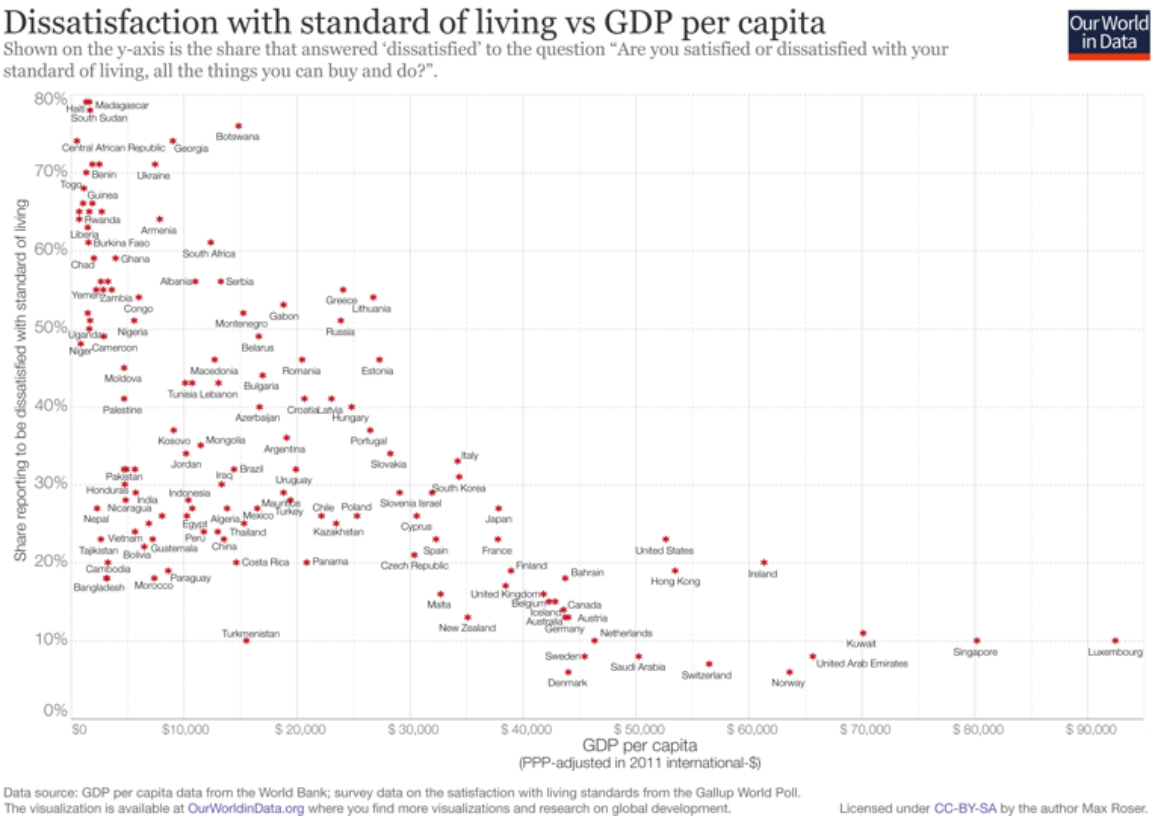
The availability heuristic can deceive you (although Kaj Sotala notes that e.g. the Bay area homeless may be benefitting less from growth than the global poor). If you just look out your window, you might miss important global trends.

Good economic policy has lifted billions of people out of poverty and furnished our lives with previously unimaginable splendor. The Roman emperors had no air conditioning or telephones.

Economics is important. And I'm unconvinced by the criticism of these numbers which I read.



Growth is important. More precisely, I argue that good economic policy → RGDP/capita growth → reduction in poverty and increase in well-being. Thus, economic policy is important to get right. Some economists seem to know how to get certain areas of economic policy right, and so I think it's worth learning from them.



Doubly amazing given hedonic treadmill & the higher expectations of people in wealthier nations.

Economists are not responsible for all of this growth. I wasn't able to quickly find counterfactual estimates for the importance of economic theories, but my impression is that several advances in economics have in fact significantly improved economic policy.

Good economic policy is a weapon against suffering, against disease and disorder and squalor, and perhaps one day against death itself.

Good economic policy makes selfish actors conspire to deliver cheap, delicious pastries to your doorstep in less than an hour.

Good economic policy is often about expanding the pie, instead of fighting over who gets what part today. This very moment, a vast universe full of energy and resources burns away.

Before reading this book, I tried some other econ textbooks. They were bad. Before the bad textbooks, economic analysis had seemed like "just another consideration." I'd taken an econ class in college, and kiiiinda remembered what "deadweight loss" meant.

When I read Cowen and Tabarrok's *Modern Principles of Economics,* I regularly felt beliefs getting debunked and replaced with less wrong beliefs—an experience last felt while reading the sequences.

As a teenager, I thought that:

- stimulus = good because obviously people need money in a recession, and they'll spend that money
- tax cuts = bad because they won't be spent as much, and usually tax cuts are just excuses to reduce tax burden on the rich

- outgroup members thought homeless people were lazy, but obviously that wasn't usually true, and those homeless people need direct fiscal help
- price gouging is bad because it's wrong to take advantage of people in emergencies

But I didn't know anything about aggregate demand, or the costs/benefits of expansionary fiscal policy, or poverty traps, or prices-as-signals. I'd basically just absorbed sentiments from my political upbringing—I recently noticed that I disliked "supply-side economics" without even knowing what that *is*! These sentiments were sometimes mostly right, and sometimes incredibly wrong.

This book brought me two benefits. First, it introduces important frames for thinking. Second, it has lots of interesting facts and compelling philosophical arguments.

So let's go.

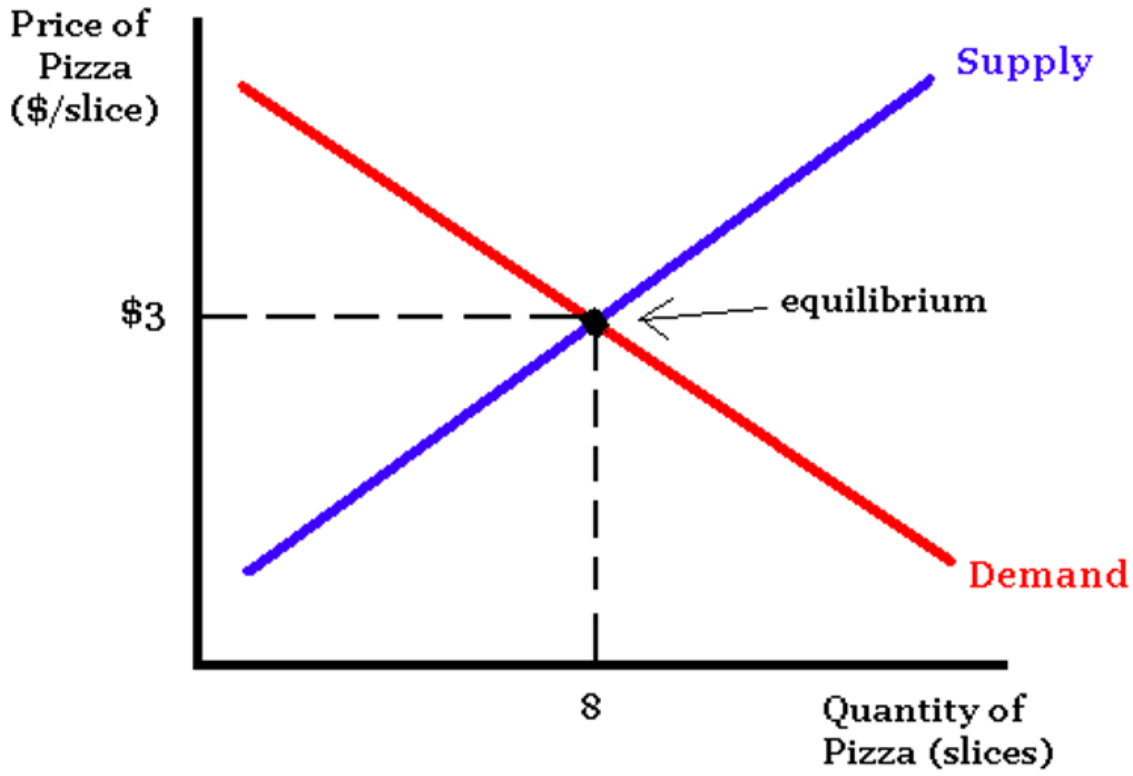
> The prisoners were dying of scurvy, typhoid fever, and smallpox, but nothing was killing them more than bad incentives. In 1787, the British government had hired sea captains to ship convicted felons to Australia. Conditions on board the ships were monstrous; some even said the conditions were worse than on slave ships. On one voyage, more than one-third of the men died and the rest arrived beaten, starved, and sick. A first mate remarked cruelly of the convicts, "Let them die and be damned, the owners have [already] been paid for their passage."
>
> The British public had no love for the convicts, but it wasn't prepared to give them a death sentence either. Newspapers editorialized in favor of better conditions, clergy appealed to the captains' sense of humanity, and legislators passed regulations requiring better food and water, light and air, and proper medical care. Yet the death rate remained shockingly high. Nothing appeared to be working until an economist suggested something new. Can you guess what the economist suggested?
>
> Instead of paying the captains for each prisoner placed on board ship in Great Britain, the economist suggested paying for each prisoner that walked off the ship in Australia. In 1793 , the new system was implemented and immediately the survival rate shot up to 99%. One astute observer explained what had happened: "Economy beat sentiment and benevolence."'
>
> The story of the convict ships illustrates the first big lesson that runs throughout this book and throughout economics:

Incentives matter.

How do people decide what to buy and where to work, what opportunities to take and where to build? *Microeconomics* models decision-making by consumers and firms. Basic microeconomic models assume that people want to make money, and they're good at it—they are rational. Unsurprisingly, this isn't quite true, but the models let us easily think about what incentives people have in different situations.

(We can add corrections to the Econ 101 arguments later. I think this is better than throwing up your hands and saying "Econ doesn't have all the answers, people are too complicated!")

The most important microeconomic frame I deeply internalized was **supply/demand curves**.



At a price of $3, suppliers will produce 8 slices of pizza.

Law of supply: Firms want to supply more pizza if you'll pay them more; the supply curve is increasing.

Law of demand: Consumers want to buy less pizza if you charge them more; the demand curve is decreasing.

(Not all markets follow these "laws.")

I significantly sharpened my understanding of incentives by internalizing how to shift supply/demand curves. So let's reason through a contentious question with this frame:

I think price gouging should usually be legal (and most economists agree)

This section serves both as an epistemic spot check and an explanation I wish I'd read when I started learning econ. Skip if it's old news to you.

Price gouging occurs when an emergency happens (e.g. a blizzard), people demand a lot of some good (e.g. snow shovels), and so stores jack up the prices (e.g. $4 -> $30).

Consider a competitive snow shovel market, where firms can price shovels as they please (or, *far* more accurately: in response to economic conditions). When demand increases for snow shovels, that's a *positive demand shock* because people want to buy more shovels. The demand curve moves out to the right, from *D* to *D'*:

Because people want more snow shovels, the price increases from P_1 to P_2. (This is the "price gouging" part.) So here is the painful part of the picture. Now snow shovels are expensive, and some people can't afford them, and also *fuck you* to the people taking advantage of a disaster just for a few bucks. Many people have this gut-level reaction.

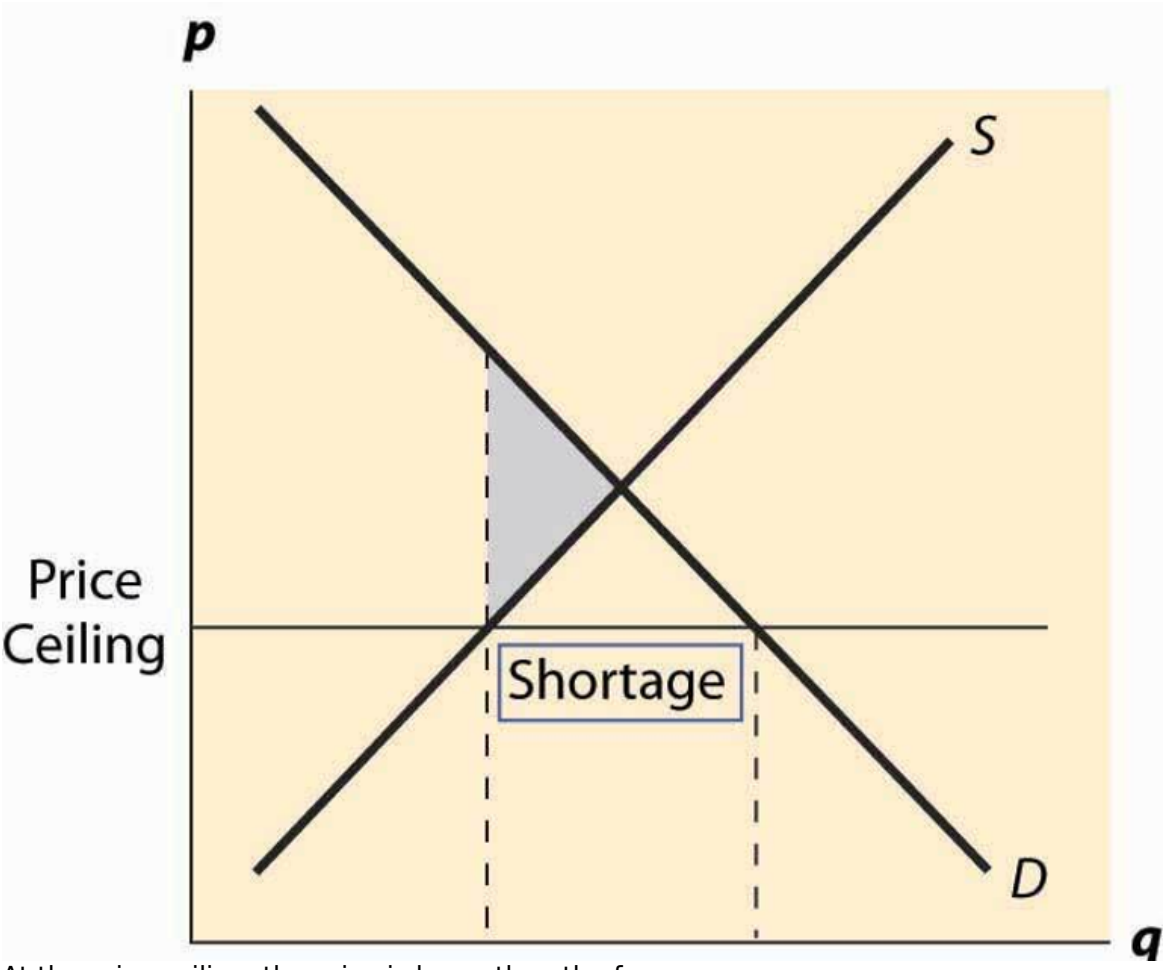
But what comes next? Suppose the storm hits Wisconsin. Demand goes up, so prices go up. *Since firms want to make money*, suppliers in neighboring states (e.g. Iowa and Illinois) will start trucking in snow shovels and selling them at a high—but slightly lower—price. In fact, the hardest-hit areas with the highest prices will get prioritized for more supply, so that firms can make more money. As more firms enter the Wisconsin snow shovel market, they compete over price and eventually the price settles back to the original P_1 as demand subsides.

Prices are signals about which places want which goods, and free markets maximize social benefit when firms make money by responding to those price signals. The high price of snow shovels is like a huge neon sign which spells **hey if you bring more snow shovels here you can make a lot of profit!**

Some laws ban price gouging. In certain industries and during an emergency, firms basically can't raise prices unless they can prove that their operating costs / input costs increased.

FN: SRAS

Suppose that the snowshovel price can barely increase from P_1 due to a so-called price ceiling.



At the price ceiling, the price is lower than the free-market equilibrium. At this artificially low price, consumers want to buy a lot of snow shovels (the second dotted line) but suppliers don't want to produce as many.

Then there's a shortage, because more people are willing to buy shovels at P_1 (what a deal, especially in an emergency!) than suppliers are willing to sell at P_1. There's *no economic incentive* for them to increase production, and incentives matter. But set aside profit-making for a moment.

The price signal—that big neon sign—is no longer present. Suppose you own a chain of shovel stores throughout the midwest USA, and all you care about is getting shovels to people who need them. Several cities all got hit by the winter storm. Which one needs the most shovels? How many?

Because there's a price ceiling, you can't tell. So you send more shovels to all the cities, but some get too many shovels and some get too few. The shortage remains, and people can't get enough snow shovels, but at least they aren't getting *cheated*, right?

In the price-gouging world, yes, some people pay more for shovels. A key issue seems to be inequality: Many poor people won't be able to afford shovels. But price-gouging has several benefits:

- Firms profit by supplying more shovels
- Firms know which areas need shovels the most, because they're the areas with highest prices
- People with the highest willingness to pay will pay higher prices, ensuring that shovels don't get misallocated to people using them for trivial reasons
- In shortages, people don't pay extra money, they pay extra time. They pay search costs as they drive around town and look online for places with shovels. In fact, the rational consumer pays up to (willingness to pay - artificially low sticker price) in search cost, erasing the supposed benefit to consumers.
    - At least a bribe has someone getting paid! Consumer time is wasted.
    - Who do you think has the ability to drop everything and run to the store to get a few shovels before they're all gone? Are they probably poorer, or probably more wealthy?
    - So it's not clear that anti-price-gouging laws even help poor people.

(Since people hate price gouging so much, large firms like Wal-Mart may decide not to raise prices and just sell out to preserve reputation, while smaller vendors gouge away. Thus you can get the "best of both worlds" without anti-price-gouging laws.)

I note that this picture assumes a competitive market; in particular, other firms can enter to sell shovels and compete to drive down prices. If that's not true, then I think that the arguments for price gouging are much weaker.

The best argument I can think of against price gouging is that people are probably more irrational and manipulable during an emergency. I don't think this overrides the other benefits of price gouging. I looked around for other counter-arguments and didn't find any I thought were good.

So that's the theory. How do things work back on planet Earth?

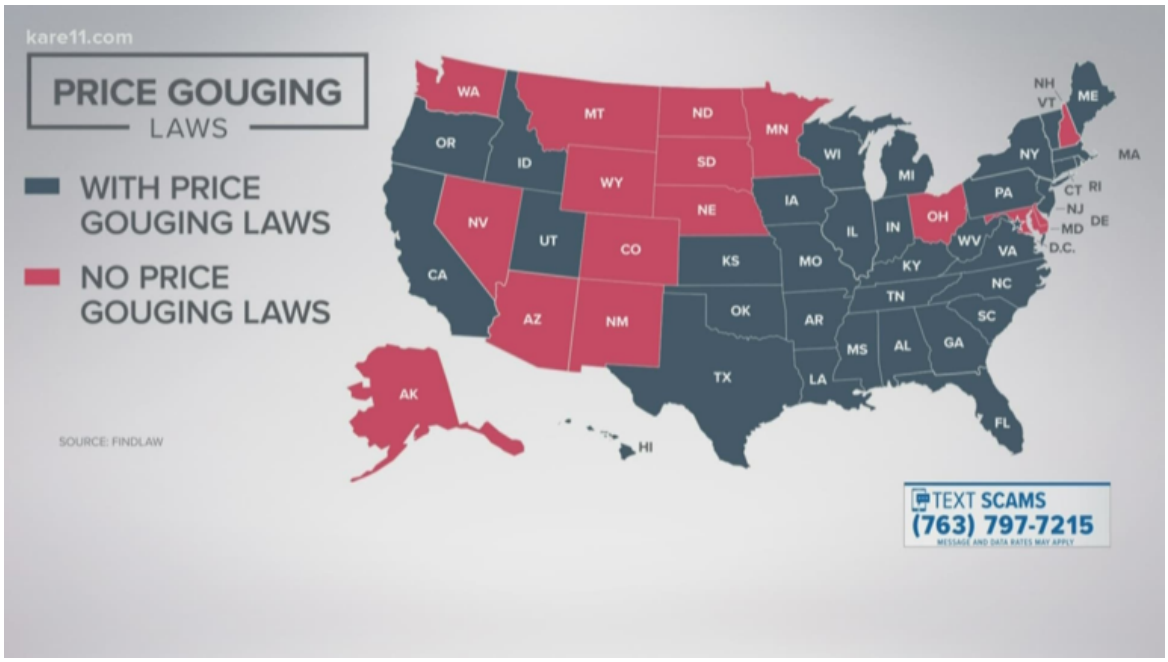
The empirical situation lines up with the theory.

> These results support standard economic theory regarding price ceilings and counter political rhetoric in support of anti-price-gouging (APG) laws. When APG laws bind, counties may also experience shortages, increases in the total price paid for goods and services due to waiting time, and adjustments on non-price margins such as quality adjustments, the development of black markets, and rationing by violence.
>
> — The Effects Of Anti-Price-Gouging Laws In The Wake Of A Hurricane

Remember the toilet paper hoarding of early 2020? APG laws are partially to thank.

The bottom line is: *In an emergency, there won't be enough shovels for everyone to get a shovel at standard price. If you want a long shortage, if you want to feel moral and avoid being blamed—outlaw price gouging in competitive markets.*

I'm sure you'll be absolutely shocked to learn that lots of states make price gouging a criminal offence.



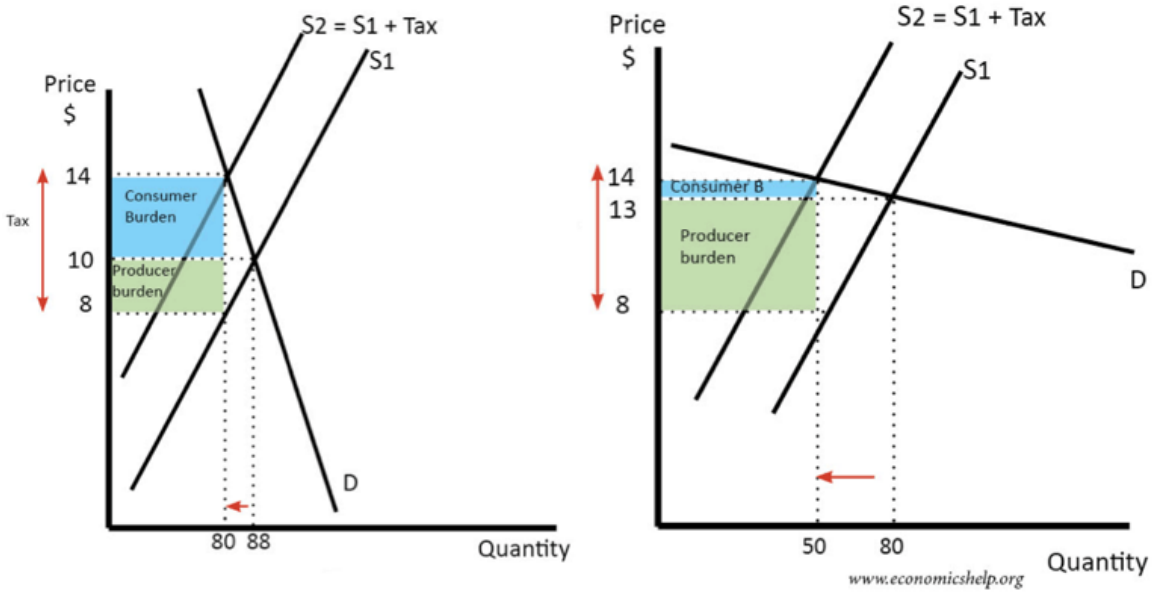
People **hate** price gouging! Notice how the non-APG states get the scary *red* color and the "scams" hotline.

How do demand shocks ripple?

By this point in the book, I've pinned down supply/demand curves. This was actually a bit tough, because I got caught up in the dynamics of how e.g. demand shocks in one industry would ripple through the economy. The answer was: Don't worry about it. That's much harder. Just focus on supply/demand curves, and everything will be OK for now.

Determining tax incidence

Another great mental motion is the "wedge" trick for tax incidence determination.



Left: Suppose that cigarette packs cost $10 at competitive equilibrium (where S1 meets D), and a $6 excise (per-item) tax is levied. A slightly more elastic supply curve means that consumers bear slightly more tax burden. **Right:** A more elastic demand curve means that producers bear most of the tax burden.

Think of the tax as the difference between what the consumer pays and the seller receives. Tax incidence answers: How much does the consumer pay, and how much does the seller receive?

The "wedge trick" is this. Consider a vertical line segment of length 6 in the above charts, and imagine it floating in from the left until it hits *S1* and *D* with its endpoints. This determines tax incidence: The supply curve shifts up to *S2*. (This is actually the same as just shifting up *S1* by $6, but I find it easier to visualize.)

If demand is more elastic than supply (aka the slope is less steep), then consumers bear more of the tax burden, and vice versa. The intuition is that it's harder to tax people who are more ready to stop consuming/producing the product.

Surprisingly, statutory incidence—who writes the check to the IRS—doesn't matter. The tax incidence is the same whether you charge consumers $6 more for buying cigarettes, or producers $6 more for selling cigarettes.

Other updates

Before I talk about the book holistically, here are more snippets:

- The much-maligned capitalism is actually probably the greatest incentive alignment success in human history.
  - This is the whole point made by the "invisible hand" idea: Competitive markets "invisibly" have each firm produce until price equals marginal cost, and this minimizes total cost to society. While each firm pursues selfish interests, they advance societal goals by creating huge amounts of value in order to make money.
  - (Just think about how much you'd be willing to pay for glasses, and how much you actually have to pay! And that market isn't even that competitive! Isn't consumer surplus amazing?)

- Rent controls are bad. Controls on competitive markets are generally bad.
    - I gained gears-level models for when government should or shouldn't be involved in a market.
- Sweatshops are bad but probably better for children than whatever else they would have been doing.
    - The Econ 101 argument goes: If the children had something better to do—like go to school or work at a safe, well-paying, age-appropriate job—they would do that instead of working at the sweatshop.
- A firm perfectly price discriminates (PPD) if they charge each consumer their willingness to pay for a good. I really hope that firms don't somehow achieve PPD via AI before the singularity. That would really suck. You'd basically be *just on the edge* of every single purchase you make—exhausting. And producers would gobble up surplus from consumers, which seems unfair.
- Custodians in the US have higher real wages than custodians in India, even though they may be equally good at cleaning. This is because American custodians are more (economically) productive since they work for *more productive* firms.
    - The marginal product of labor (MPL) is the revenue from hiring an additional worker.
    - A custodian at Google is adding more value than an equally custodian at a small Indian firm.
    - Therefore, they have higher MPL, and are paid more.
        - Concretely: American custodians average $30,906/year. Adjusting for Indian purchasing power parity, that's equivalent to $30,906 × $\frac{₹22}{\$1}$ = 679,932 rupees. The average Indian custodian only makes ₹120,726/year—less than a fifth the American wage.
    - Your wages aren't just determined by your skills and your work ethic, but by the rest of the economy. It pays to work in a wealthy economy!
- I had looked forward to learning about fractional reserve banking, ever since HJPEV reflected on the crudity of the Gringotts banking system. "What a sophisticated-sounding economic idea", I thought. "I'll need to study carefully to understand that one day", I thought.
    - I thought wrong. Fractional reserve banking is stupidly simple. Banks keep a fraction of deposits on hand as "reserves", in case customers want to make withdrawals. The rest of the deposits are loaned out. These loans put the money to work, growing the economy with people's savings while also ensuring that people can withdraw their money.
        - This is why "just lock your gold underground" Gringotts is dumb: The wizarding economy could be growing using some of Lucius Malfoy's money. Lucius, Gringotts, and the economy would benefit from this arrangement.
    - This is a general pattern: The key concepts in economics do not require nearly as much mental scaffolding as math concepts do. (Try explaining topological continuity in three sentences.)
        - *Applying* economic reasoning is still a delicate endeavor, though.
- I now know what the Federal Reserve is and what they do.
    - And I now appreciate how hard central banking can be.
    - And I now appreciate how great it is that we let *actual economists* run this part of the government. Even if they sometimes mess up.
    - And I now appreciate this huge body of literature on interest rates and quantitative easing and M1 and market monetarism and *agh*! I'd always consigned the finance section of the paper to "inscrutable garbage that daytraders worry about", but it's so much *more*. I can't wait to keep learning about macroeconomics.
- The fact that wages are sticky-down is so annoying. I would be so mad if I were a central banker: dumb human bias ruining our ability to deflate! Dumb human bias making negative AD shocks horrible!
    - The dumb bias seems to be nominal wage confusion—responding to the dollar amount on your paycheck (the nominal wage), instead of to the goods you can

buy with the money (the real wage).
    - In a negative aggregate demand shock, nominal GDP growth decreases. Normally firms could just lower wages for all their employees and real output would remain the same. But people hate hate hate seeing nominal wage cuts, and so it's easier for firms to just fire people, or at least raise real wages far more slowly than inflation.
    - Which cuts actual growth in the short run.
    - What a mess.

Reflections

- I'm very glad that I've used Anki. I probably made over 500 cards for this book, not only for the key vocab but also for charts, for brain-teasers, for cool pieces of reasoning the authors used. Cloze deletions are fast and convenient for all of these purposes.
- This book is long. 800 pages long. It covers both micro- and macro-economics, and I was pleasantly surprised by the macro part. I'd heard macro is garbage, but I think it's just less understood than micro.
- investopedia and econlib are great resources for learning about economics.
- It took a while for me to get comfortable with economics. At first, I felt uncomfortable and reluctant to read more, because everything seemed mildly confusing. Now I can read papers (with great effort) and have a good idea what they're talking about. Learning more is now easy and fun; I've crossed the hump for economics in the same way I crossed the hump for mathematics.
- I'm glad I read this book. It's long, and maybe I could have skipped some of it. I didn't get much out of the advanced indifference curve chapter; it wasn't presented clearly. Most of the book was quite clear.

Conclusion

Economics is interesting, and powerful. Famine used to haunt most of the world, and now it doesn't.

> Famine seems to be the last, the most dreadful resource of nature. The power of population is so superior to the power of the earth to produce subsistence for man, that premature death must in some shape or other visit the human race. The vices of mankind are active and able ministers of depopulation. They are the precursors in the great army of destruction, and often finish the dreadful work themselves. But should they fail in this war of extermination, sickly seasons, epidemics, pestilence, and plague advance in terrific array, and sweep off their thousands and tens of thousands. Should success be still incomplete, gigantic inevitable famine stalks in the rear, and with one mighty blow levels the population with the food of the world.
>
> — Thomas Malthus, 1798. *An Essay on the Principle of Population*.

Appendix: Open questions I have

I haven't seriously looked into these questions yet.

- What's going on with market monetarism and quantitative easing?
    - In particular, I still find open market operations confusing for some reason.
- Why isn't the risk-free rate subtracted from capital returns?
    - I suspect this is just government being dumb.
- Why aren't all contracts indexed to official inflation estimates?

    - Wouldn't this significantly cut down on arbitrary wealth transfers from inflation/deflation?
    - TIPS is indexed. What else?
    - This can't be explained by dumb government; contracts are private.
- Why do firms in a cartel have individual incentive to raise price above competitive?
    - IE: Why isn't competitive pricing a Nash? If one firm unilaterally raised prices, wouldn't the other firms just sell more in their wake?
- What is the shadow banking system?
- Study the impossible trinity of international economics.
- How does free banking work?

FN: SRAS. Talk about some awful upwards-sticky prices; I wonder what the effect on SRAS is? I'd guess that ceilings probably aren't binding for long enough to show up much macroeconomically.

Thanks to LessWrong for feedback on this post.

Do a cost-benefit analysis of your technology usage

If an unaligned entity invests billions of dollars into an application which you use, where they benefit from wasting your time, and you haven't at least done a cost-benefit analysis so that your usage minimizes your costs and maximizes your benefits—*You are probably getting fucked over*.

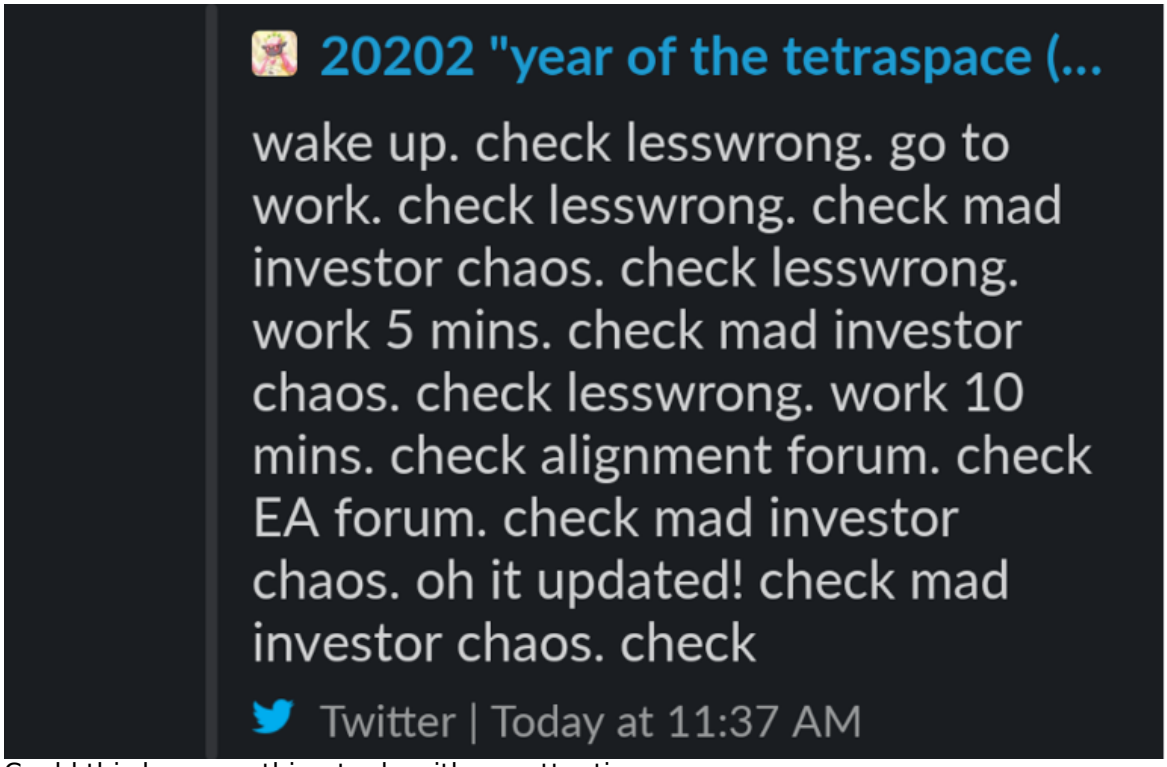
Mistake: Motivatedly avoiding thinking about the issue

Last summer, my friend Kurt Brown told me about *Digital Minimalism*. The modern world is mired in attention-sucking apps which compete to waste as much of your time as possible. The book's remedy: stepping back from non-essential internet usage, so that you can evaluate what really matters to you. After a month has come and gone, you add back in those digital activities which are worth it to you.

Unfortunately, this is the part of the story where we all cringe at my past behavior. I gave Kurt some excuses, demurring from his implicit recommendation that I read the book. I asked more questions, but so that I could learn more about what he'd been up to. I wasn't going to actually do it. I think it sounded monastic and uncomfortable and *I'm not one of those people who needs it, I already have lots of locks on my devices*.

And locks I had. I restricted my iPhone with a password only known by a friend, so that I was unable to access *eg* Reddit without wiping my device, or asking my friend for the code. My phone was in black and white to minimize how appealing it would be, I had an outdated model to make using my phone less enjoyable. I didn't have notifications for anything but phone calls. I *still wasted several hours a day on my phone, although I was always (motivatedly) surprised by this.* I thought I was spending at least 70% of my phone-time productively, by reading LessWrong and Wikipedia, or engaging in work communication. In this scenario, I didn't want to upend my life for a month in order to save less than an hour a day (even though it still would have been worth it in the long run).

This school year, I've had problems focusing and relaxing. I tried exercise, different medication, but nothing hit the spot. I wasn't reading textbooks like I wanted to, my attention was fractured, I often felt behind my schedule. I was still doing my job and making progress—just not as much as I wanted.



Could this have anything to do with my attention problems?

This spring, I read a LessWrong post which mentioned *Digital Minimalism*. Luckily, this triggered my "if several reasonably smart EAs swear by the benefits of X, investigate X" trigger-action plan.

Digital Minimalism

I listened to the first half of the book on Audible in one night. As I wrote above:

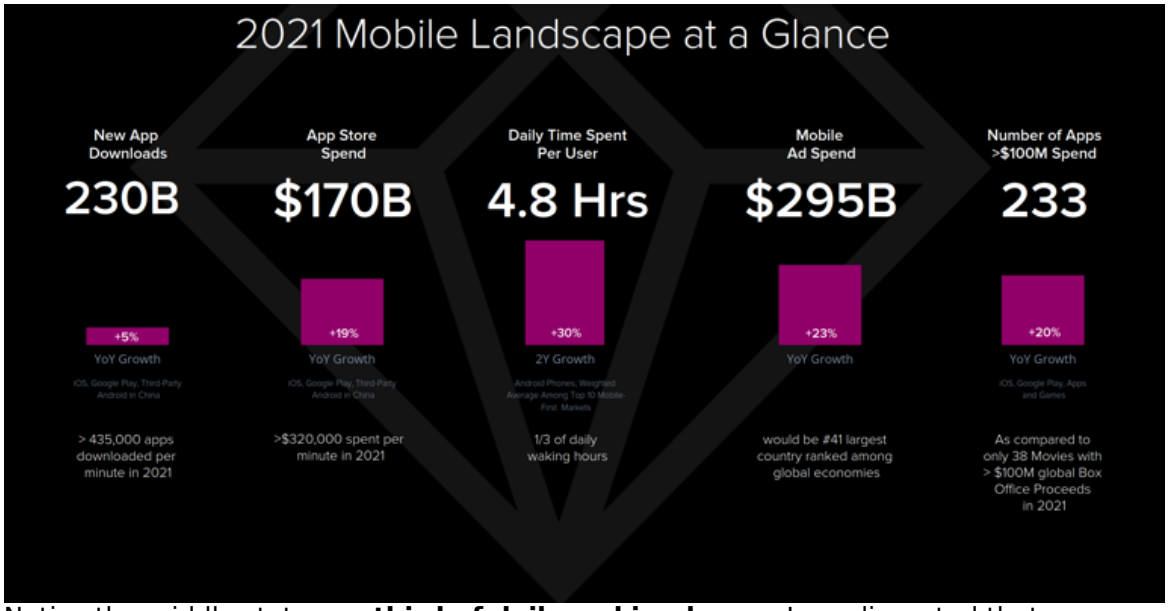
> If an unaligned entity invests billions of dollars into an application which you use, where they benefit from wasting your time, and you haven't at least done a cost-benefit analysis so that your usage minimizes your costs and maximizes your benefits—*You are probably getting fucked over*.[1]

I was immediately convinced that this thesis is correct, and resolved to start my month-long "digital declutter" the next day.

Time costs

Consider why you originally bought a cell phone. It was probably to call people, to text people, to take photos, to get GPS navigation. Would you have bought it if you foresaw how you would feel an urge to check it even during a dinner with a friend you hadn't seen in a long time? Would you have bought it if you knew you would take impatiently take it out of your pocket dozens of times a day, staring at it 2+ hours daily?

The point isn't "phone bad, never use phone, quit now." My phone provides me with enormous benefits. The point is *where was the cost-benefit analysis, what tf has happened to us?!*



Notice the middle stat: **one third of daily waking hours**. I am disgusted that some people try to make this number go up further. From AppAnnie.

Readers of this forum are probably better about their usage. Let's be (too) generous and cut that to a mere two hours wasted daily on your phone, and 0 hours wasted on your other devices. That's only *one eighth of your waking year,* or 1.5 waking months each year.

Attentional costs

But lost time doesn't capture everything sucked away by your apps, by your email tics, by YouTube, by Reddit, by Slack, by Discord, by everything else which is after you. *Digital Minimalism* asked:

> When was the last time you were bored and in silence?

I remember lazy summer childhoods, staring at the ceiling after I ran out of video game time. At my 2018 CFAR workshop, my phone dipped in a stream for several minutes and short-circuited. I was actually glad. I felt free. How strange, to feel *free* from a device I purchased! Perhaps I should have noticed the warning sign.

Since then, engagement has been a pocket-grasp away. I'd leave my phone in another room to work, only to find my way back half an hour later. Even now, I look down at my phone on my desk, and I feel it. I feel it calling to me from far away, whispering to me, urging me to check Slack or my email—just one more time.

These compulsions kill deep work in the cradle. My attention was fractured and strewn. I would anxiously procrastinate by flitting through tabs: *Discord. Slack. LessWrong. Gmail.* Even when I cleared time to think, I would periodically check my phone.

Implementing the declutter

At this point, you might be thinking "OK, but I can't roam the mountains of Nepal for a month. I have work to do and that requires staying in touch with people." Sure. The point of this post is not "no phone." The point of this post is to build a digital life purposefully and carefully, because you reflectively endorse each component. The point of this post is to get

people to do *any cost-benefit analysis at all* of the way they spend 1/8th–1/3rd of their waking hours.

My estimate of the daily costs and benefits for a better-than-average Facebook user (considering Messenger to be distinct from Facebook). In appendix 2, I detail how I extract

all of these benefits for 40 minutes a month, instead of 40 minutes a day—a 30x improvement!

The declutter goes as follows:

1. Identify the minimal set of digital affordances required to do your job and the other necessities of life (e.g. paying bills).
2. Cut out everything else for one month.

The point is that these apps which are out to get you—they're very good at what they do. It's not enough to turn off notifications and enable app timers. *Digital Minimalism* argues (and I mostly agree) that you have to *get out of the pond entirely* and catch a breather. After the declutter, you can soberly analyze the costs and benefits of each digital activity you add back in.

My declutter rules

I went by a whitelist[2] in order to ensure there wasn't a way to weasel around the rules. Here's what I let myself do:

- Phone
  - Voice & video calls
  - GPS
  - Audible
  - Uber/Lyft
  - Authenticators/alarms/other boring utilities
  - Roam/note-taking
- iPad
  - Note-taking
  - Reading
  - Drawing
- Computer
  - Anything offline (except music or video games)
  - Textbooks and Wikipedia and arxiv/google scholar
  - Overleaf for writing papers
  - Amazon and Upwork (for managing contracted-out labor)
  - Zoom for weekly meetings
  - Anki and Roam
  - Check email at noon on Mondays and Thursdays
    - I told people to call me if it was important. I didn't get any calls.
    - I later let myself send emails without looking at my inbox. I recommend Inbox When Ready, which hides your inbox by default and prevents you from being attention-sniped.
  - Groceries / other mundane things
    - No Wealthfront—no reason for me to see how my portfolio is doing.

That's it. No music (see appendix), no messaging, no Facebook, no Twitter, no Slack, no Discord, no anxious email checking, no Youtube, no nothing. I even bought a cheap watch so that I wouldn't have to check my phone for the time. If I needed an exception, I'd first write a note explaining what I did, to be read by my girlfriend, Emma, who started her declutter soon after.

Why did I choose these rules? I won't get unhealthily sucked into any of these activities. They all make me stronger. They let me do my work.

The world was not going to end because I stopped reading the news for a month—I understand there's a war still going on in Ukraine, but that's about all I know, and I'm not

worse off for it. I resolved that if I wanted to build models of that part of the world, I'd do that on purpose. I won't doomscroll through hyper-optimized interfaces designed to scam me out of my attention and make me anxious. I said to myself, my life is worth more to me than that.

FAQ

> But TurnTrout, my job needs email / [other special reason why this doesn't work for me].

I concede that my rules are probably not best for your situation. But have you thought about the issue for five minutes? Could you ask your boss if you can check email once a day and otherwise take phone calls? Maybe you don't restrict email, but stop looking at websites like Reddit or Hacker News or Marginal Revolution or Facebook or Twitter? Are there other creative solutions waiting to be uncovered? Have you tried?

If your team uses Slack for asynchronous communication, once- or twice-daily checks should be fine. If you use it for synchronous communication, perhaps establish a daily "office hour" when you'll be on Slack, or even coordinate with your team to establish a daily "Slack hour" where people are expected to be online. Or something else. The point is to establish the main benefits you reap from each digital affordance, and then find a plan which minimizes the costs you pay for those benefits.

> I'm already good about my internet usage.

This might be true! I know exactly one person for whom I'm quite confident this is true (Andrew Critch), and maybe there are more among my friends whom I don't know about. This might be you if you already use services based on their costs and benefits, often using websites in unintended ways (like blocking all recommendations on YouTube via the Unhook add-on [Chrome, Firefox]), and spending far less time than average (eg only checking email very infrequently).

I'd still bet against it. I would have said I was good about my internet usage, and it *was* true—in a relative sense. I think people (motivatedly) underestimate how much time they waste, perhaps because it can feel bad and embarrassing to admit the problem.

> But how will I stay in touch with people? I'm already lonely.

Excellent question! Reallocate low-quality social time to high-quality social time. Instead of checking if some half-friends liked your FB status, call up a buddy and grab a beer, or go to a meetup, or join a club.

Benefits of the declutter

February 22nd: The First Day. I went running, and got back to the house 10 minutes earlier than usual. Huh.

I called my parents and went on a leisurely walk. Even so, I got my morning routine done 60 minutes ahead of schedule. I read half of a book on ordinary differential equations while lounging in my sunlit room. I did some deep thinking for an hour, safe from my phone's dopaminergic temptation. I switched contexts and read about electrostatics. Still hours ahead of schedule.

The day yawned and stretched. I wondered if it would ever end. (It did.)

February 23rd: The Second Day. From my journal:

> It's so relaxing not using my phone, and yet I can still feel my anxiety pulling me to my digital affairs.

> Did my LW post get lots of upvotes? Are people criticizing me? Did I win a prize in the contest? Am I missing something on the EliezerFic server? I even thought about some identity-politics tweet I saw last week, on my run this morning... why is that garbage in my head? Good riddance.

And so unrolled the next day, and the next. Time laid itself out before me. With my reclaimed time, I went on walks, I read *The Character of Physical Law*, I read ~three physics textbooks, I tripled my daily Anki workload to 1.5 hours, and I *still had time left over*.

Life became leisurely. I wrote letters to my girlfriends—some of them were in French. I even had time to write *poems*. I talked to them more often than before, with nightly phone calls. I also called my family most mornings. *I still had time left over*.

Instead of trolling through Discord, I called some labmates at Oregon State and started a weekly dinner. If anything, I felt less lonely than before, when I had the world at my fingertips. I called people when I wanted to talk to them. *I still had time left over.*

I listened to a Stephen King book when I couldn't sleep—I found it reassuring to worry about fortifying a grocery store against eldritch horrors, instead of worrying about fortifying our planet against artificial intelligence. I listened to Dune with Emma, clocking 21 hours over 2.5 weeks. I went on walks with her, and to a hot tub, and I *still had time left over*.

I did notice potential withdrawal symptoms (alarming!), mostly via increased baseline anxiety. Other explanations include "defending my dissertation & moving soon", so I'm not sure if it was from the declutter.

Even assuming this month gave me unusually large benefits, I wouldn't ever, ever go back. So when the declutter ended, I wasn't clamoring to check the highest-karma Reddit posts from last month. I still feel the urge, but I resent Reddit now that I see what it takes away from me. That makes it easier to stay away.

Recommendations

This short post may not be convincing enough to try out such a substantial life modification. I'm not asking that you do a declutter right away. I'm recommending that you read the first half of *Digital Minimalism*, or listen on Audible (cost: a few hours and $14).

Let me sweeten the deal with a costly signal. If I've met you in real life, and you consume the first half of the book and find it unconvincing / try the declutter and it wasn't at all worth trying in hindsight, message me on LessWrong and I'll pay you $30.[3]

I think many, many people are shooting themselves in the foot, so I will be blunt. *Please stop shooting yourself in the foot.* Please do a cost-benefit analysis. I think many people have serious, serious problems with their internet usage. I did. You might. If so, you are leaving a lot of your life on the table.

Thanks to Meg Tong, Josh Turner, and Kurt Brown for feedback on this post.

Appendix 1: Declutter advice

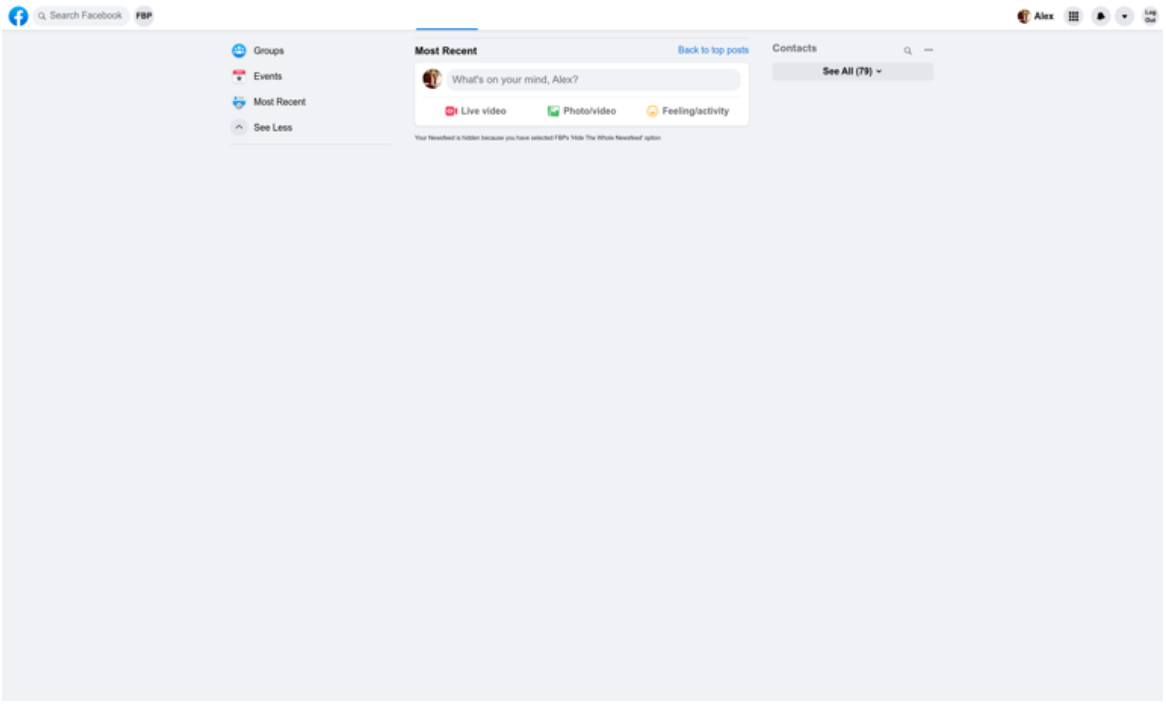
Here's my main tip to add to the book: *Have well-defined exception handling which you never ever ever have to deviate from.* When I read about how other people navigated the declutter, their main failure modes looked like "my dog died and I got really stressed and gave in" or "a work emergency came up and I bent my rules and then broke my rules fragrantly."

Plan for these events. Plan for feeling withdrawal symptoms. Plan for it seeming *so so important* that you check your email *right now*. Plan for emergencies. Plan a way to handle surprising exceptions to your rules. Make the exception handling so good that you never have a legitimate reason to deviate from it.

My procedure was “If I need to use a forbidden functionality, then I have to write what I did down on a slip of paper and leave it on my girlfriend’s desk ASAP.” This worked because Emma would understand legitimate exceptions, but would look askance at me if I started flooding her desk with “and then I checked Reddit” notes. It's easier to hold promises to other people, than promises to yourself.

Appendix 2: My post-declutter rules

- I only listen to music when:
  - Only listening to the music, to fully soak it in
  - Exercising
  - Reasoning on this point:
    - I think music generally makes me subtly dumber but feel cooler while I’m listening to it, so I listen to it a lot.
    - Music imposes its own form on my thoughts. My thinking and mood becomes governed by the song which happens to be playing, and less by the substance of my own thoughts. I don’t want my reasoning to hinge on “will Spotify shuffle to *Attack on Titan* or *Coldplay* next?”.
    - See also Gwern’s stub.
    - I do have Google Home, and often play nature sounds.
- I only check LessWrong / Discord / Slack / Messenger / my text messages each Sunday at noon.
  - I write blogposts before then, and I won’t check their reception until the next week (I used to nervously refresh).
  - I’ve also adblocked the karma elements of the website, because I worry too much about them.
- As I currently see it, I’m only logging in to the newsfeed part of Facebook two more times: To share this blog post, and after I receive my PhD.
  - After that, I’ll check its event page weekly, while blocking the notifications / other clutter FB tries to throw at me. This should take less than 10 minutes each week.
  - Here’s how to use FB more peacefully:
    - Install FBPurity; you can save time by importing my settings here.
    - Use UBlock Origin to get rid of the rest; here is my element blocking list for Facebook.
      - (I also hide the chat sidebar on the main page, which is a FB option)
  - I could also check a favorite page once a week (with the chat and comment elements blocked), if I need more memes in my bloodstream.
  - In combination with a monthly Messenger checkin, I’ve extracted my main benefits from Facebook, at a cost of at most 50 minutes each month, instead of 50 minutes each day!
    - Again, I ***don’t recommend*** doing small fixes like “just hide some FB elements.” These fixes **don’t work** for most people. This advice is aimed at post-declutter usage, which unfolds from your informed cost-benefit analysis.



Here's what my FB news feed looks like now. 😊

- For news, I purchased a digital+print subscription to ~~*The Economist*~~. Once a month, I can choose to read the four issues for an hour or two.
    - I don't need to read more than that. I can read about candidates before an election, and there isn't much else that's decision-relevant. If eg AI dynamics heat up and geopolitical understanding becomes more important, I'll tackle that deliberately.
    - Looking back at my life, I see how often I've been hijacked by news websites. It makes me sick.
    - UPDATE: No longer recommend *The Economist.* Their cancellation process is scummy, recommend avoiding the defectors.
- I'm basically not going to text anymore. I used to check it so, so often.
    - This was hard at first. One of my partners strongly prefers texting, and I liked texting her, and missed her a lot. With additional thought, we discovered that she really just wanted to asynchronously send me updates on how her day was going. I said she could text me as much as she wanted—but I'd read them during our next phone call.
- I can watch movies and play video games if I've planned it out at least a few hours in advance.
- I can check Reddit for specific question/answer threads.
- I can check Twitter if I plan the session out in detail one day in advance.
    - Twitter is toxic for me, even though I originally made an account to promote an alignment paper and only subscribed to AI/math accounts.
- My phone will still be in black and white and warm color temperature, to make it even less engaging compared to the rest of my world.
- I *never ever use my phone on the toilet. Ever.* This has served me well and seems like a pure win.

1. ^

   This is only a sufficient condition; the app need not be the child of a billion-dollar company. For example, I oft ragebaited myself about the culture war via *Marginal*

Revolution and *Hacker News*. I even tend to get anxious about *LessWrong* usage, and I know that the team deliberately refrained from attention-hacks like red notifications.

Even while using my Notion to edit this post and supervise research, I saw a red “5 notifications” marker, which gave me an overwhelming urge to *see what the notifications are*. With great effort, I ignored the impulse, and deleted the element with my adblocker.

2. ^

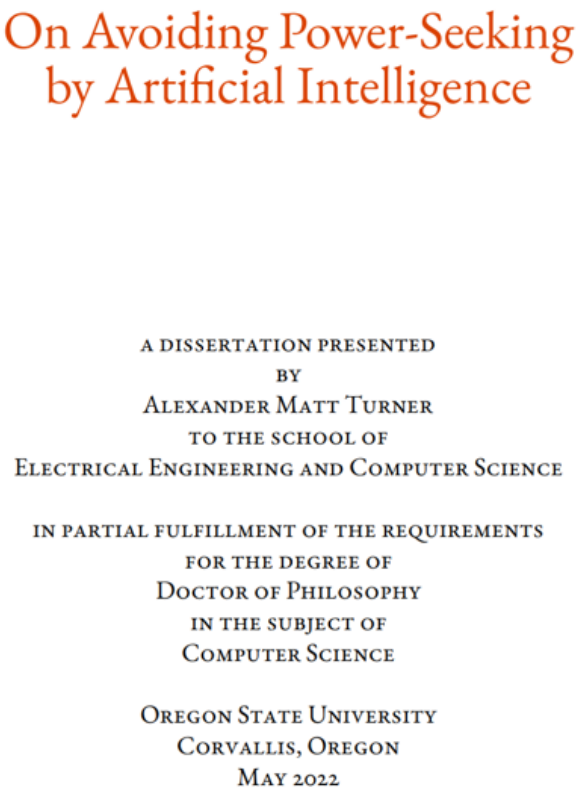
   I just now picked up my phone and stared at it blankly. One month later. Yuck.

3. ^

   Limit $300 total.

Looking back on my alignment PhD

This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on Spotify , Apple Podcasts , Libsyn , and more.


On Avoiding Power-Seeking
by Artificial Intelligence

A DISSERTATION PRESENTED
BY
ALEXANDER MATT TURNER
TO THE SCHOOL OF
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE

OREGON STATE UNIVERSITY
CORVALLIS, OREGON
MAY 2022

> My dissertation. It's long, so if you're going to read anything from it, read Chapter 0 (Introduction).

The funny thing about long periods of time is that they do, eventually, come to an end. I'm proud of what I accomplished during my PhD. That said, I'm going to first focus on mistakes I've made over the past four[1] years.

Mistakes

I think I got significantly smarter in 2018–2019, and kept learning some in 2020–2021. I *was* significantly less of a fool in 2021 than I was in 2017. That is important and worth feeling good about. But all things considered, I still made a lot of profound mistakes over the course of my PhD.

Social dynamics distracted me from my core mission

I focused on "catching up" to other thinkers

I figured this point out by summer 2021.

I wanted to be more like Eliezer Yudkowsky and Buck Shlegeris and Paul Christiano. They know lots of facts and laws about lots of areas (e.g. general relativity and thermodynamics and information theory). I focused on building up dependencies (like analysis and geometry and topology) not only because I wanted to know the answers, but because I felt I owed a *debt* , that I was *in the red* until I could at least meet other thinkers at their level of knowledge.

But rationality is not about the bag of facts you know, nor is it about the concepts you have internalized. Rationality is about *how* your mind holds itself, it is *how* you weigh evidence, it is *how* you decide where to look next when puzzling out a new area.

If I had been more honest with myself, I could have nipped the "catching up with other thinkers" mistake in 2018. I could have removed the bad mental habits using certain introspective techniques ; or at least been aware of the badness.

But I did not, in part because the truth was uncomfortable. If I did not have a clear set of prerequisites (e.g. analysis and topology and game theory) to work on, I would not have a clear and immediate direction of improvement. I would have felt adrift.

But there is not yet any "rationality tech tree", no succession of well-defined rationality skills such that you can learn them in order and grow way stronger. Like, you can't just do the calibration exercises, and then the noticing-confusion exercises, and then other things. Those tools help, but they aren't *enough*. There won't be a clear and immediate direction of improvement, at first. But you may want to get stronger anyways.

I focused on seeming smart and defensible

I figured this point out this spring .

When I started working on alignment, I didn't know what to do at first, and I felt insecure about my credentials. As far as I remember, I figured I'd start off by becoming respected, since other people's feedback was initially a better guide than my own taste. Unfortunately, I didn't realize how deeply and subtly this goal would grow its roots.

I worried about upvotes, I worried about winning arguments, I worried about being defensible against criticism. I was so worried that someone would comment on one of my posts and tear everything down, because I *hadn't been careful enough*, because I had *left myself open* by not dotting all my 'i's. (Not that anyone has ever done that on LessWrong before...)

I think it was this year that I had my (second) "oh man, *don't forget the part where everyone is allowed to die to AI*" moment. To illustrate the new mindset this gut-realization gave me, I'll detail a recent decision with social consequences, and then compare the old and the new mindsets.

A few months back, Quintin Pope approached me with (what he claimed to be) a new alignment paradigm, which blossomed from asking the following kind of questions:

> We clearly prefer future AIs to generalize in the way that neuroscientists generalize, so it seems worthwhile to ask: "why don't neuroscientists wirehead themselves?"
>
> It's clearly not because humans evolved away from wireheading, *specifically*. There are somewhat similar situations to wireheading in the ancestral environment: psychoactive drugs, masturbation, etc. Is the reason we don't wirehead because evolution instilled us with an aversion to manipulating our reward function, which then zero-shot generalized to wireheading, despite wireheading being so wildly dissimilar to the contents of the

> ancestral environment? How could evolution have developed an alignment approach that generalized so well?

After a few days, I realized my gut expectations were that he was broadly correct and that this theory of alignment could actually be right. However, I realized I wasn't consciously letting myself think that because it would be Insufficiently Skeptical to actually think the alignment problem is solvable. This seemed obviously stupid to me, so I quickly shut that line of thinking down and second-order updated towards optimism so that I would stop *predictably* getting more optimistic about Quintin's theory.[2]

I realized I assigned about 5% credence to "this line of thinking marks a direct and reasonably short path to solving alignment." Thus, on any calculation of benefits and harms, I should be willing to stake some reputation to quickly get more eyeballs on the theory, even though I expected to end up looking a little silly (with about 95% probability). With my new attitude, I decided "whatever, let's just get on with it and stop wasting time."

The old "don't leave any avenue of being criticized!" attitude would have been less loyal to my true beliefs: "This *could* work, but there are so many parts I don't understand yet. If I figure those parts out first, I can explain it better and avoid having to go out on a limb in the process." Cowardice and social anxiety, dressed up as prudence and skepticism.

I still get anxious around disagreements with people I respect. I am still working on fully expunging the "defensibility" urges, because they suck. But I've already made a lot of progress.[3]

Too much deference, too little thinking for myself

I realized and started fixing this mistake this spring . (Seeing a pattern?)

I filtered the world through a status lens. If I read a comment from a high-status person, I would gloss over confusing parts, because *I* was probably the one reading it wrong. Sure, I would verbally agree that modest epistemology is unproductive. I just *happened* to not think thoughts like "[high-status person]'s claim seems obviously dumb and wrong."

Now I let myself think thoughts like that, and it's great. For example, last week I was reading about Pavlov's conditioning experiments with dogs. I read the following:

> Pavlov (1902) started from the idea that there are some things that a dog does not need to learn. For example, dogs don’t learn to salivate whenever they see food. This reflex is ‘hard-wired’ into the dog.

I thought, "that seems like bullshit. Really, the dogs are *hard-wired* to salivate when they ***see*** food? Doesn't that require *hard-wiring a food-classifier into the dog's brain*?!"

And you know what? It *was* bullshit. I searched for about 8 minutes before finding references of the original lectures Pavlov gave:

> Dr. Zitovich took several young puppies away from their mother and fed them for considerable time only on milk. When the puppies were a few months old he established fistulae of their salivary ducts, and was thus able to measure accurately the secretory activity of the glands. **He now showed these puppies some solid food -- bread or meat -- but no secretion of saliva was evoked**.

Our world is so inadequate that seminal psychology experiments are described in mangled, misleading ways. Inadequacy abounds, and status only weakly tracks adequacy. Even if the high-status person belongs to your in-group. Even if all your smart friends are nodding along.

Would you notice if *this very post* were inadequate and misleading? *Would* it be bullshit for the dog-genome to hardwire a food-classifier? *Think for yourself. Constant vigilance!*

Non-social mistakes

I thought about comfortable, familiar problems

I figured this point out this spring, because I bumped into Quintin as described above.

I remember a sunny summer day in 2019, sitting in the grass with Daniel Filan at UC Berkeley. He recommended putting together an end-to-end picture of the alignment problem. I remember feeling pretty uncomfortable about that, feeling that I wouldn't understand which alignment problems go where in my diagram ("do embedded agency failures crop up *here*, or *there*?"). Wouldn't it just make more sense to read more alignment papers and naturally refine those views over time?

This was a rationalization, plain and simple. There is no point where you feel ready to put all the pieces together. If you feel totally comfortable about how alignment fits together such that Daniel's exercise does not *push you* on some level, we have either *already* solved the alignment problem, or you are deluded.

I did not feel ready, and I was not ready, and I should have done it anyways. But I focused on more comfortable work with well-defined boundaries, because it felt good to knock out new theorems. Whether or not those theorems were useful and important to alignment, that was a mistake. So I stayed in my alignment comfort zone. I should have stopped working on impact measures and power-seeking way earlier than I did, even though I did end up doing some cool work.

Not admitting to myself that I thought alignment was doomed

Figured this out this spring. I'm not sure if I've fixed the general error yet.

After I became more optimistic about alignment due to having a sharper understanding of the overall problem and of how human values formed to begin with, I also became more pessimistic about *other* approaches, like IDA/ELK/RRM/AUP/[anything else with a three-letter acronym]. But my new understanding didn't seem to present any *specific* objections. So why did I suddenly feel worse about these older ideas?

I *suspect* that part of the explanation is: I hadn't wanted to admit how confused I was about alignment, and I (implicitly) clutched to "but it *could* work"-style hopefulness. But now that I had a *different* reason to hope, resting upon a more solid and mechanistic understanding, *now* it was apparently emotionally safe for me to admit I didn't have much hope at all for the older approaches.

Yikes.

If that's what happened, I was seriously deluding myself. I will do better next time.

I viewed my life through narratives

I probably figured this point out in 2021.

Back in 2018, I had the "upstart alignment researcher" narrative—starting off bright-eyed and earnest, learning a lot, making friends. But then I hurt my hands and couldn't type anymore, which broke the narrative. I felt dejected—to slightly exaggerate, I felt I had fallen off of the sunlit path, and now nothing was going to go as it should.

Another example of narrative-thinking is when people say "I'm just not a math person." This is an *inference* and a *story* they tell themselves. Strictly speaking, they may not know much math, and they may not enjoy math, and they may not see how to change either of those facts. But the *narrative* is that they are not a math person. Their discomfort and their aversion-to-trying stem not just from their best-guess assessment of their own weaknesses, but from a *story* they are living in.

Every moment is an opportunity for newly-directed action. Keep your identity small and keep the narratives in the story-books. At least, if you want to use narratives, carefully introspect to make sure you're using them, and they aren't using you.

Other helpful habits I picked up

I'm not really sure where these two habits go, so I'll put them here. I wish I'd had these skills in 2018.

- **Distinguish between *observations* and *inferences*.** When people speak to you, mark their arguments as *observations* or as *inferences*. Keep the types *separate*. I've gained *so much* from this simple practice.

  Here are two cases I've recently found where people seem to mistake the folk wisdom for observation:
  - "People often say they're afraid to die" is an *observation,* and "people are hard-wired to be afraid of death" is an *inference.*
  - "I often feel 'curiosity' and some kind of exploration-impulse" is an *observation,* and "people are innately curious" is an *inference.*
- **Be concrete.** My friend Kurt remarks that I constantly ask for examples.
  - If a friend comes to me for advice and says "I'm terrible at dating, I just feel so shy!", I *could* say "You're really fun to be around, you're probably just in your head too much", and then *they* could say "Agh, maybe, but it's just so frustrating." Wouldn't that just be such a useful conversation for them? That'll *definitely* solve their awkwardness!
    - Alternatively, if I *ask for an example*, we can both analyze an event which *actually happened*. Perhaps they say, "I met a girl named Alice at the party, but I somehow ran out of things to say, and it got quiet, and we found excuses to part ways." Then I can help my friend introspect and figure out why they didn't have anything to say, which *is in fact a question with a real answer*.
  - The general rhythm is: Bind your thinking to *coherent scenarios* (preferably ones which *actually happened*, like meeting a girl named Alice), so that you (and possibly other people) can explore the details together (like why it got quiet) in order to figure out what to change (like running mock encounters to shoo away the social anxiety).
  - On the other hand, if you can't think of a concrete example to ground your airy words, maybe your thinking is totally untethered from reality. Maybe your assumptions are *contradictory* and you can't even see it.
    - Here's something I recently said on Discord:

      "If there are some circuits who can defer to the market prediction, then each circuit can get their coalitional contribution as their fixed weight. This lets some relatively simpler circuits retain weight. At least, those are the abstract words I want to say, but now I feel confused about how to apply that to a concrete example for how e.g. a shallow but broad "don't steal" value negotiates via Critch-bargaining. **Not being able to give a concrete example means I don't really know what I'm talking about here.**"

- Don't tell me how your alignment strategy will e.g. "faithfully reproduce human judgments." Explain what concrete benefits you hope to realize, and why "faithful reproduction of human judgments" will realize those benefits.
  - If the actual answer is that you *don't know*, then just *say it*, because it's the truth. Be aware that you don't know.

To close out the "Mistakes" section, I mostly wish I'd expected more from myself. I wish I'd believed myself capable of building an end-to-end picture of the alignment problem, of admitting what I didn't know and what I hadn't thought about, of being able to survive/ignore the harsh winds of criticism and skepticism.

I did these things eventually, though, and I'm proud of that.

What I'm proud of

1. I didn't keep working on computational chemistry. Boy howdy, would that have been awful for me. *Thank you,* TurnTrout_{2018}!
    1. I remember thinking "You know what, I'd rather get *expelled* than not do [the 2018 CHAI internship]." This thought gave me the courage to find a new advisor who would let me work on AI safety, funding be damned.
    2. I'm not a natural nonconformist. Conflict makes me nervous. I've had to work for it.
2. I learned a lot of math, even though I felt sheepish and insecure about it at first.
3. I think I ended up achieving rationality escape velocity.
    1. When I get stuck / feel depressed, errors get thrown, exception-handling activates, I start thinking "these thoughts seem unreasonably dark; my cognition is compromised; have I eaten enough food today, have I drank enough water, should I call a friend...".
    2. When I get stuck on a problem (e.g. what is the type signature of human values?), I do not stay stuck. I notice I am stuck, I run down a list of tactics, I explicitly note what works, I upweight that for next time.
    3. When I realize I've been an idiot about something (e.g. nicking my hand with a knife, missing a deadline), I stop and think *wow, that was stupid, what's the more general error I'm making*?
    4. The general rhythm is: I feel agentic and capable and self-improving, and these traits are strengthening over time, as is the rate of strengthening.
    5. This definitely didn't have to happen, but I made it happen (with the help of some friends and resources).
4. Research achievements:
    1. I think *Reframing Impact* correctly inferred our intuitions around what "impact" means, and also that sequence was beautiful and I loved making it.
    2. My dissertation is also beautiful. I painstakingly wrote and formatted and edited it, even hiring a professional to help out. I fought to keep its tone focused on what matters: the sharp dangers of AGI.
    3. I likewise poured myself into *Optimal Policies Tend To Seek Power*, and its follow-up, *Parametrically Retargetable Decision-Makers Tend To Seek Power*.
        1. First, I had felt instrumental convergence should be provable and formally understandable. It was a mystery to me in 2019, and now it's not.
        2. Second, I used to suck at writing academic papers, but I managed to get two NeurIPS spotlights by the end of my program. NeurIPS spotlights might not save the world, but that was tough and I did a good job with it.
    4. Attainable utility preservation is pointless for AGI alignment, but *damn* is it cool that we could do unsupervised learning to get a reward function, preserve the agent's ability to optimize that single random objective, and *just get cautious behavior in complicated environments*.

Looking forward

Leaving Oregon was a bit sad, but coming to Berkeley is exciting. I'll be starting my CHAI postdoc soon. I'm working with lots of cool, smart, loyal friends. I'm feeling strong and confident and relatively optimistic, both about alignment and about my personal future.

Here's to winning. 🥂

1. ^

   My PhD was six years long (it started in the fall of 2016). However, I'm not even going to critique the first two years, because that would make the "Mistakes" section far too long.

2. ^

   If you're interested in reading about the theory now, see this recent comment. I'm currently putting together some prerequisite posts to bridge the inferential gap.

3. ^

   Sometimes I feel the urge to defend myself *just a little more*, to which some part of me internally replies "are you serious, this defensibility thing again?! Are you *ever* going to let me *actually think*?"

   I like that part of me a lot.