Best of LessWrong: May 2019

1. Yes Requires the Possibility of No
2. Offer of collaboration and/or mentorship
3. Coherent decisions imply consistent utilities
4. Integrating disagreeing subagents
5. Separation of Concerns
6. Naked mole-rats: A case study in biological weirdness
7. Complex Behavior from Simple (Sub)Agents
8. Say Wrong Things
9. Quotes from Moral Mazes
10. Interpretations of "probability"
11. Which scientific discovery was most ahead of its time?
12. Where are people thinking and talking about global coordination for AI safety?
13. Authoritarian Empiricism
14. Probability interpretations: Examples
15. Towards optimal play as Villager in a mixed game
16. Ed Boyden on the State of Science
17. Announcing my YouTube channel
18. Totalitarian ethical systems
19. Nash equilibriums can be arbitrarily bad
20. Natural Structures and Conditional Definitions
21. More Notes on Simple Rules
22. Physical linguistics
23. Correspondence visualizations for different interpretations of "probability"
24. Space colonization: what can we definitely do and how do we know that?
25. What is required to run a psychology study?
26. 345M version GPT-2 released
27. What makes a scientific fact 'ripe for discovery'?
28. Evidence for Connection Theory
29. And the AI would have got away with it too, if...
30. A shift in arguments for AI risk
31. Von Neumann’s critique of automata theory and logic in computer science
32. Kevin Simler's "Going Critical"
33. Episode 3 of Tsuyoku Naritai! (the 'becoming stronger podcast): Nike Timers
34. "One Man's Modus Ponens Is Another Man's Modus Tollens"
35. Eight Books To Read
36. Blame games
37. Why exactly is the song 'Baby Shark' so catchy?
38. A quick map of consciousness
39. Simple Rules of Law
40. By default, avoid ambiguous distant situations
41. How much do major foundations grant per hour of staff time?
42. April 2019 gwern.net newsletter
43. Schelling Fences versus Marginal Thinking
44. Episode 5 of Tsuyoku Naritai (the 'becoming stronger' podcast): The Stance
45. Does the Higgs-boson exist?
46. Programming Languages For AI
47. Is AI safety doomed in the long term?
48. alternative history: what if Bayes rule had never been discovered?
49. Dishonest Update Reporting
50. Hierarchy and wings

Best of LessWrong: May 2019

1. Yes Requires the Possibility of No
2. Offer of collaboration and/or mentorship
3. Coherent decisions imply consistent utilities
4. Integrating disagreeing subagents
5. Separation of Concerns
6. Naked mole-rats: A case study in biological weirdness
7. Complex Behavior from Simple (Sub)Agents
8. Say Wrong Things
9. Quotes from Moral Mazes
10. Interpretations of "probability"
11. Which scientific discovery was most ahead of its time?
12. Where are people thinking and talking about global coordination for AI safety?
13. Authoritarian Empiricism
14. Probability interpretations: Examples
15. Towards optimal play as Villager in a mixed game
16. Ed Boyden on the State of Science
17. Announcing my YouTube channel
18. Totalitarian ethical systems
19. Nash equilibriums can be arbitrarily bad
20. Natural Structures and Conditional Definitions
21. More Notes on Simple Rules
22. Physical linguistics
23. Correspondence visualizations for different interpretations of "probability"
24. Space colonization: what can we definitely do and how do we know that?
25. What is required to run a psychology study?
26. 345M version GPT-2 released
27. What makes a scientific fact 'ripe for discovery'?
28. Evidence for Connection Theory
29. And the AI would have got away with it too, if...
30. A shift in arguments for AI risk
31. Von Neumann’s critique of automata theory and logic in computer science
32. Kevin Simler's "Going Critical"
33. Episode 3 of Tsuyoku Naritai! (the 'becoming stronger podcast): Nike Timers
34. "One Man's Modus Ponens Is Another Man's Modus Tollens"
35. Eight Books To Read
36. Blame games
37. Why exactly is the song 'Baby Shark' so catchy?
38. A quick map of consciousness
39. Simple Rules of Law
40. By default, avoid ambiguous distant situations
41. How much do major foundations grant per hour of staff time?
42. April 2019 gwern.net newsletter
43. Schelling Fences versus Marginal Thinking
44. Episode 5 of Tsuyoku Naritai (the 'becoming stronger' podcast): The Stance
45. Does the Higgs-boson exist?
46. Programming Languages For AI
47. Is AI safety doomed in the long term?
48. alternative history: what if Bayes rule had never been discovered?
49. Dishonest Update Reporting

50. Hierarchy and wings

Yes Requires the Possibility of No

1. A group wants to try an activity that really requires a lot of group buy in. The activity will not work as well if there is doubt that everyone really wants to do it. They establish common knowledge of the need for buy in. They then have a group conversation in which several people make comments about how great the activity is and how much they want to do it. Everyone wants to do the activity, but is aware that if they did not want to do the activity, it would be awkward to admit. They do the activity. It goes poorly.

2. Alice strongly wants to believe A. She searches for evidence of A. She implements a biased search, ignoring evidence against A. She finds justifications for her conclusion. She can then point to the justifications, and tell herself that A is true. However, there is always this nagging thought in the back of her mind that maybe A is false. She never fully believes A as strongly as she would have believed it if she just implemented an an unbiased search, and found out that A was, in fact, true.

3. Bob wants Charlie to do a task for him. Bob phrases the request in a way that makes Charlie afraid to refuse. Charlie agrees to do the task. Charlie would have been happy to do the task otherwise, but now Charlie does the task while feeling resentful towards Bob for violating his consent.

4. Derek has an accomplishment. Others often talk about how great the accomplishment is. Derek has imposter syndrome and is unable to fully believe that the accomplishment is good. Part of this is due to a desire to appear humble, but part of it stems from Derek's lack of self trust. Derek can see lots of pressures to believe that the accomplishment is good. Derek does not understand exactly how he thinks, and so is concerned that there might be a significant bias that could cause him to falsely conclude that the accomplishment is better than it is. Because of this he does not fully trust his inside view which says the accomplishment is good.

5. Eve is has an aversion to doing B. She wants to eliminate this aversion. She tries to do an internal double crux with herself. She identifies a rational part of herself who can obviously see that it is good to do B. She identifies another part of herself that is afraid of B. The rational part thinks the other part is stupid and can't imagine being convinced that B is bad. The IDC fails, and Eve continues to have an aversion to B and internal conflict.

6. Frank's job or relationship is largely dependent to his belief in C. Frank really wants to have true beliefs, and so tries to figure out what is true. He mostly concludes that C is true, but has lingering doubts. He is unsure if he would have been able to conclude C is false under all the external pressure.

7. George gets a lot of social benefits out of believing D. He believes D with probability 80%, and this is enough for the social benefits. He considers searching for evidence of D. He thinks searching for evidence will likely increase the probability to 90%, but it has a small probability of decreasing the probability to 10%. He values the social benefit quite a bit, and chooses not to search for evidence because he is afraid of the risk.

8. Harry sees lots of studies that conclude E. However, Harry also believes there is a systematic bias that makes studies that conclude E more likely to be published, accepted, and shared. Harry doubts E.

9. A bayesian wants to increase his probability of proposition F, and is afraid of decreasing the probability. Every time he tries to find a way to increase his probability, he runs into an immovable wall called the conservation of expected evidence. In order to increase his probability of F, he must risk decreasing it.

Offer of collaboration and/or mentorship

UPDATE: Offer of mentorship is closed, since I received sufficiently many candidates for now. Offer of collaboration remains open for experienced researchers (i.e. researchers that (i) have some track record of original math / theoretical compsci research, and (ii) are able to take on concrete open problems without much guidance).

I have two motivations for making this offer. First, there have been discussions regarding the lack of mentorship in the AI alignment community, and that beginners find it difficult to enter the field since the experienced researchers are too busy working on their research to provide guidance. Second, I have my own research programme which has a significant number of shovel ready open problems and only one person working on it (me). The way I see it, my research programme is a very promising approach that attacks the very core of the AI alignment problem.

Therefore, I am looking for people who would like to either receive mentorship in AI alignment relevant topics from me, or collaborate with me on my research programme, or both.

Mentorship

I am planning to allocate about 4 hours / week to mentorship, which can be done over Skype, Discord, email or any other means of remote communication. For people who happen to be located in Israel, we can do in person sessions. The mathematical topics in which I feel qualified to provide guidance include: linear algebra, calculus, functional analysis, probability theory, game theory, computability theory, computational complexity theory, statistical/computational learning theory. I am also more or less familiar with the state of the art in the various approaches other people pursue to AI alignment.

Naturally, people who are interested in working on my own research programme are those who would benefit the most from my guidance. People who want to work on empirical ML approaches (which seem to be dominant in OpenAI, DeepMind and CHAI) would benefit somewhat from my guidance, since many theoretical insights from computational learning theory in general and my own research in particular, are to some extent applicable even to deep learning algorithms whose theoretical understanding is far from complete. People who want to work on MIRI's core research agenda would also benefit somewhat from my guidance but I am less knowledgeable or interested in formal logic and approaches based on formal logic.

Collaboration

People who want to collaborate on problems within the learning-theoretic research programme might receive a significantly larger fraction of my time, depending on details. The communication would still be mostly remote (unless the collaborator is in Israel), but physical meetings involving flights are also an option.

The original essay about the learning-theoretic programme does mention a number of more or less concrete research directions, but since then more shovel ready problems joined the list (and also, there are a couple of new results). Interested people are advised to contact me to hear about those problems and discuss the details.

Contact

Anyone who wants to contact me regarding the above should email me at vanessa.kosoy@intelligence.org, and give me a brief intro about emself, including knowledge in math / theoretical compsci and previous research if relevant. Conversely, you are welcome to browse my writing on this forum to form an impression of my abilities. If we find each other mutually compatible, we will discuss further details.

Coherent decisions imply consistent utilities

(*Written for Arbital in 2017.*)

Introduction to the introduction: Why expected utility?

So we're talking about how to make good decisions, or the idea of 'bounded rationality', or what sufficiently advanced Artificial Intelligences might be like; and somebody starts dragging up the concepts of 'expected utility' or 'utility functions'.

And before we even ask what those are, we might first ask, *Why?*

There's a mathematical formalism, 'expected utility', that some people invented to talk about making decisions. This formalism is very academically popular, and appears in all the textbooks.

But so what? Why is that *necessarily* the best way of making decisions under every kind of circumstance? Why would an Artificial Intelligence care what's academically popular? Maybe there's some better way of thinking about rational agency? Heck, why is this formalism popular in the first place?

We can ask the same kinds of questions about probability theory:

Okay, we have this mathematical formalism in which the chance that X happens, aka P(X), plus

the chance that X doesn't happen, aka P(¬X), must be represented in a way that makes the two

quantities sum to unity: P(X) + P(¬X) = 1.

That formalism for probability has some neat mathematical properties. But so what? Why should the best way of reasoning about a messy, uncertain world have neat properties? Why shouldn't an agent reason about 'how likely is that' using something completely unlike probabilities? How do you *know* a sufficiently advanced Artificial Intelligence would reason in probabilities? You haven't seen an AI, so what do you think you know and how do you think you know it?

That entirely reasonable question is what this introduction tries to answer. There are, indeed, excellent reasons beyond academic habit and mathematical convenience for why we would by default invoke 'expected utility' and 'probability theory' to think about good human decisions, talk about rational agency, or reason about sufficiently advanced AIs.

The broad form of the answer seems easier to show than to tell, so we'll just plunge straight in.

Why not circular preferences?

De gustibus non est disputandum, goes the proverb; matters of taste cannot be disputed. If I like onions on my pizza and you like pineapple, it's not that one of us is right and one of us is wrong. We just prefer different pizza toppings.

Well, but suppose I declare to you that I *simultaneously*:

- Prefer onions to pineapple on my pizza.
- Prefer pineapple to mushrooms on my pizza.
- Prefer mushrooms to onions on my pizza.

If we use $>_P$ to denote my pizza preferences, with $X >_P Y$ denoting that I prefer X to Y, then I am declaring:

$$\text{onions} >_P \text{pineapple} >_P \text{mushrooms} >_P \text{onions}$$

That sounds strange, to be sure. But is there anything *wrong* with that? Can we disputandum it?

We used the math symbol $>$ which denotes an ordering. If we ask whether $>_P$ can be an ordering, it naughtily violates the standard transitivity axiom $x > y,\ y > z \implies x > z$.

Okay, so then maybe we shouldn't have used the symbol $>_P$ or called it an ordering. Why is that necessarily bad?

We can try to imagine each pizza as having a numerical score denoting how much I like it. In that case, there's no way we could assign consistent numbers x, y, z to those three pizza toppings such that $x > y > z > x$.

So maybe I don't assign numbers to my pizza. Why is that so awful?

Are there any grounds besides "we like a certain mathematical formalism and your choices don't fit into our math," on which to criticize my three simultaneous preferences?

(Feel free to try to answer this yourself before continuing...)

Click here to reveal and continue:

Suppose I tell you that I prefer pineapple to mushrooms on my pizza. Suppose you're about to give me a slice of mushroom pizza; but by paying one penny ($0.01) I can instead get a slice of pineapple pizza (which is just as fresh from the oven). It seems realistic to say that most people with a pineapple pizza preference would probably pay the penny, if they happened to have a penny in their pocket.[1]

After I pay the penny, though, and just before I'm about to get the pineapple pizza, you offer me a slice of onion pizza instead—no charge for the change! If I was telling the truth about preferring onion pizza to pineapple, I should certainly accept the substitution if it's free.

And then to round out the day, you offer me a mushroom pizza instead of the onion pizza, and again, since I prefer mushrooms to onions, I accept the swap.

I end up with exactly the same slice of mushroom pizza I started with... and one penny poorer, because I previously paid $0.01 to swap mushrooms for pineapple.

This seems like a *qualitatively* bad behavior on my part. By virtue of my incoherent preferences which cannot be given a consistent ordering, I have shot myself in the foot, done something self-defeating. We haven't said *how* I ought to sort out my inconsistent preferences. But no matter how it shakes out, it seems like there must be *some* better alternative—some better way I could reason that wouldn't spend a penny to go in circles. That is, I could at least have kept my original pizza slice and not spent the penny.

In a phrase you're going to keep hearing, I have executed a 'dominated strategy': there exists some other strategy that does strictly better.[2]

Or as Steve Omohundro put it: If you prefer being in Berkeley to being in San Francisco; prefer being in San Jose to being in Berkeley; and prefer being in San Francisco to being in San Jose; then you're going to waste a lot of time on taxi rides.

None of this reasoning has told us that a non-self-defeating agent must prefer Berkeley to San Francisco or vice versa. There are at least six possible consistent orderings over pizza toppings, like mushroom $>_P$ pineapple $>_P$ onion etcetera, and *any* consistent ordering would avoid paying to go in circles.[3] We have not, in this argument, used pure logic to derive that pineapple pizza must taste better than mushroom pizza to an ideal rational agent. But we've seen that eliminating a certain kind of shoot-yourself-in-the-foot behavior, corresponds to imposing a certain *coherence* or *consistency* requirement on whatever preferences are there.

It turns out that this is just one instance of a large family of *coherence theorems* which all end up pointing at the same set of core properties. All roads lead to Rome, and all the roads say, "If you are not shooting yourself in the foot in sense X, we can view you as having coherence property Y."

There are some caveats to this general idea.

For example: In complicated problems, perfect coherence is usually impossible to compute—it's just too expensive to consider *all* the possibilities.

But there are also caveats to the caveats! For example, it may be that if there's a powerful machine intelligence that is not *visibly to us humans* shooting itself in the foot in way X, then *from our perspective* it must look like the AI has coherence property Y. If there's some sense in which the machine intelligence is going in circles, because *not* going in circles is too hard to compute, well, *we* won't see that either with our tiny human brains. In which case it may make sense, from our perspective, to think about the machine intelligence *as if* it has some coherent preference ordering.

We are not going to go through all the coherence theorems in this introduction. They form a very large family; some of them are a *lot* more mathematically intimidating; and honestly I don't know even 5% of the variants.

But we can hopefully walk through enough coherence theorems to at least start to see the reasoning behind, "Why expected utility?" And, because the two are a package deal, "Why probability?"

Human lives, mere dollars, and coherent trades

An experiment in 2000—from a paper titled "The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals"—asked subjects to consider the dilemma of a hospital administrator named Robert:

> Robert can save the life of Johnny, a five year old who needs a liver transplant, but the transplant procedure will cost the hospital $1,000,000 that could be spent in other ways, such as purchasing better equipment and enhancing salaries to recruit talented doctors to the hospital. Johnny is very ill and has been on the waiting list for a transplant but because of the shortage of local organ donors, obtaining a liver will be expensive. Robert could save Johnny's life, or he could use the $1,000,000 for other hospital needs.

The main experimental result was that most subjects got angry at Robert for even considering the question.

After all, you can't put a dollar value on a human life, right?

But better hospital equipment also saves lives, or at least one hopes so.[4] It's not like the other potential use of the money saves zero lives.

Let's say that Robert has a total budget of $100,000,000 and is faced with a long list of options such as these:

- $100,000 for a new dialysis machine, which will save 3 lives
- $1,000,000 for a liver for Johnny, which will save 1 life
- $10,000 to train the nurses on proper hygiene when inserting central lines, which will save an expected 100 lives
- ...

Now suppose—this is a supposition we'll need for our theorem—that Robert *does not care at all about money,* not even a tiny bit. Robert *only* cares about maximizing the total number of lives saved. Furthermore, we suppose for now that Robert cares about every human life equally.

If Robert does save as many lives as possible, given his bounded money, then Robert must *behave like* somebody assigning some consistent dollar value to saving a human life.

We should be able to look down the long list of options that Robert took and didn't take, and say, e.g., "Oh, Robert took all the options that saved more than 1 life per $500,000 and rejected all options that saved less than 1 life per $500,000; so Robert's behavior is *consistent* with his spending $500,000 per life."

Alternatively, if we can't view Robert's behavior as being coherent in this sense—if we cannot make up *any* dollar value of a human life, such that Robert's choices are consistent with that dollar value—then it must be possible to move around the same amount of money, in a way that saves more lives.

We start from the qualitative criterion, "Robert must save as many lives as possible; it shouldn't be possible to move around the same money to save more lives." We end up with the quantitative coherence theorem, "It must be possible to view Robert as trading dollars for lives at a consistent price."

We haven't proven that dollars have some intrinsic worth that trades off against the intrinsic worth of a human life. By hypothesis, Robert doesn't care about money at all. It's just that every dollar has an *opportunity cost* in lives it could have saved if deployed differently; and this opportunity cost is the same for every dollar because money is fungible.

An important caveat to this theorem is that there may be, e.g., an option that saves a hundred thousand lives for $200,000,000. But Robert only has $100,000,000 to spend. In this case, Robert may fail to take that option even though it saves 1 life per $2,000. It was a good option, but Robert didn't have enough money in the bank to afford it. This does mess up the elegance of being able to say, "Robert must have taken *all* the options saving at least 1 life per $500,000", and instead we can only say this with respect to options that are in some sense small enough or granular enough.

Similarly, if an option costs $5,000,000 to save 15 lives, but Robert only has $4,000,000 left over after taking all his other best opportunities, Robert's last selected option might be to save 8 lives for $4,000,000 instead. This again messes up the elegance of the reasoning, but Robert is still doing exactly what an agent *would* do if it consistently valued lives at 1 life per $500,000—it would buy all the best options *it could afford* that purchased at least that many lives per dollar. So that part of the theorem's conclusion still holds.

Another caveat is that we haven't proven that there's some specific dollar value in Robert's head, as a matter of psychology. We've only proven that Robert's outward behavior can be *viewed as if* it prices lives at *some* consistent value, assuming Robert saves as many lives as possible.

It could be that Robert accepts every option that spends less than $500,000/life and rejects every option that spends over $600,000, and there aren't any available options in the middle. Then

Robert's behavior can equally be *viewed as* consistent with a price of $510,000 or a price of $590,000. This helps show that we haven't proven anything about Robert explicitly *thinking* of some number. Maybe Robert never lets himself think of a specific threshold value, because it would be taboo to assign a dollar value to human life; and instead Robert just fiddles the choices until he can't see how to save any more lives.

We naturally have not proved by pure logic that Robert must want, in the first place, to save as many lives as possible. Even if Robert is a good person, this doesn't follow. Maybe Robert values a 10-year-old's life at 5 times the value of a 70-year-old's life, so that Robert will sacrifice five grandparents to save one 10-year-old. A lot of people would see that as entirely consistent with valuing human life in general.

Let's consider that last idea more thoroughly. If Robert considers a preteen equally valuable with 5 grandparents, so that Robert will shift $100,000 from saving 8 old people to saving 2 children, then we can no longer say that Robert wants to save as many 'lives' as possible. That last decision would decrease by 6 the total number of 'lives' saved. So we can no longer say that there's a qualitative criterion, 'Save as many lives as possible', that produces the quantitative coherence requirement, 'trade dollars for lives at a consistent rate'.

Does this mean that coherence might as well go out the window, so far as Robert's behavior is concerned? Anything goes, now? Just spend money wherever?

"Hm," you might think. "But... if Robert trades 8 old people for 2 children *here*... and then trades 1 child for 2 old people *there*..."

To reduce distraction, let's make this problem be about apples and oranges instead. Suppose:

- Alice starts with 8 apples and 1 orange.
- Then Alice trades 8 apples for 2 oranges.
- Then Alice trades away 1 orange for 2 apples.
- Finally, Alice trades another orange for 3 apples.

Then in this example, Alice is using a strategy that's *strictly dominated* across all categories of fruit. Alice ends up with 5 apples and one orange, but could've ended with 8 apples and one orange (by not making any trades at all). Regardless of the *relative* value of apples and oranges, Alice's strategy is doing *qualitatively* worse than another possible strategy, if apples have any positive value to her at all.

So the fact that Alice can't be viewed as having any coherent relative value for apples and oranges, corresponds to her ending up with qualitatively less of some category of fruit (without any corresponding gains elsewhere).

This remains true if we introduce more kinds of fruit into the problem. Let's say the set of fruits Alice can trade includes {apples, oranges, strawberries, plums}. If we can't look at Alice's trades and make up some relative quantitative values of fruit, such that Alice could be trading consistently with respect to those values, then Alice's trading strategy must have been dominated by some other strategy that would have ended up with strictly more fruit across all categories.

In other words, we need to be able to look at Alice's trades, and say something like:

"Maybe Alice values an orange at 2 apples, a strawberry at 0.1 apples, and a plum at 0.5 apples. That would explain why Alice was willing to trade 4 strawberries for a plum, but not willing to trade 40 strawberries for an orange and an apple."

And if we *can't* say this, then there must be some way to rearrange Alice's trades and get *strictly more fruit across all categories* in the sense that, e.g., we end with the same number of plums and apples, but one more orange and two more strawberries. This is a bad thing if Alice *qualitatively* values fruit from each category—prefers having more fruit to less fruit, ceteris paribus, for each category of fruit.

Now let's shift our attention back to Robert the hospital administrator. *Either* we can view Robert as consistently assigning some *relative* value of life for 10-year-olds vs. 70-year-olds, *or* there

must be a way to rearrange Robert's expenditures to save either strictly more 10-year-olds or strictly more 70-year-olds. The same logic applies if we add 50-year-olds to the mix. We must be able to say something like, "Robert is consistently behaving as if a 50-year-old is worth a third of a ten-year-old". If we *can't* say that, Robert must be behaving in a way that pointlessly discards some saveable lives in some category.

Or perhaps Robert is behaving in a way which implies that 10-year-old girls are worth more than 10-year-old boys. But then the relative values of those subclasses of 10-year-olds need to be viewable as consistent; or else Robert must be qualitatively failing to save one more 10-year-old boy than could've been saved otherwise.

If you can denominate apples in oranges, and price oranges in plums, and trade off plums for strawberries, all at consistent rates... then you might as well take it one step further, and factor out an abstract unit for ease of notation.

Let's call this unit *1 utilon,* and denote it €1. (As we'll see later, the letters 'EU' are appropriate here.)

If we say that apples are worth €1, oranges are worth €2, and plums are worth €0.5, then this tells us the relative value of apples, oranges, and plums. Conversely, if we *can* assign consistent relative values to apples, oranges, and plums, then we can factor out an abstract unit at will—for example, by arbitrarily declaring apples to be worth €100 and then calculating everything else's price in apples.

Have we proven by pure logic that all apples have the same utility? Of course not; you can prefer some particular apples to other particular apples. But when you're done saying which things you qualitatively prefer to which other things, if you go around making tradeoffs in a way that can be *viewed as* not qualitatively leaving behind some things you said you wanted, we can *view you* as assigning coherent quantitative utilities to everything you want.

And that's one coherence theorem—among others—that can be seen as motivating the concept of *utility* in decision theory.

Utility isn't a solid thing, a separate thing. We could multiply all the utilities by two, and that would correspond to the same outward behaviors. It's meaningless to ask how much utility you scored at the end of your life, because we could subtract a million or add a million to that quantity while leaving everything else conceptually the same.

You could pick anything you valued—say, the joy of watching a cat chase a laser pointer for 10 seconds—and denominate everything relative to that, without needing any concept of an extra abstract 'utility'. So (just to be extremely clear about this point) we have not proven that there is a separate thing 'utility' that you should be pursuing instead of everything else you wanted in life.

The coherence theorem says nothing about which things to value more than others, or how much to value them relative to other things. It doesn't say whether you should value your happiness more than someone else's happiness, any more than the notion of a consistent preference ordering $>_P$ tells us whether onions $>_P$ pineapple.

(The notion that we should assign equal value to all human lives, or equal value to all sentient lives, or equal value to all Quality-Adjusted Life Years, is *utilitarianism.* Which is, sorry about the confusion, a whole 'nother separate different philosophy.)

The conceptual gizmo that maps thingies to utilities—the whatchamacallit that takes in a fruit and spits out a utility—is called a 'utility function'. Again, this isn't a separate thing that's written on a stone tablet. If we multiply a utility function by 9.2, that's conceptually the same utility function because it's consistent with the same set of behaviors.

But in general: If we can sensibly view any agent as doing as well as qualitatively possible at *anything*, we must be able to view the agent's behavior as consistent with there being some coherent relative quantities of wantedness for all the thingies it's trying to optimize.

Probabilities and expected utility

We've so far made no mention of *probability.* But the way that probabilities and utilities interact, is where we start to see the full structure of *expected utility* spotlighted by all the coherence theorems.

The basic notion in expected utility is that some choices present us with uncertain outcomes.

For example, I come to you and say: "Give me 1 apple, and I'll flip a coin; if the coin lands heads, I'll give you 1 orange; if the coin comes up tails, I'll give you 3 plums." Suppose you relatively value fruits as described earlier: 2 apples / orange and 0.5 apples / plum. Then *either* possible outcome gives you something that's worth more to you than 1 apple. Turning down a so-called 'gamble' like that... why, it'd be a dominated strategy.

In general, the notion of 'expected utility' says that we assign certain quantities called *probabilities* to each possible outcome. In the example above, we might assign a 'probability' of 0.5 to the coin landing heads (1 orange), and a 'probability' of 0.5 to the coin landing tails (3 plums). Then the total value of the 'gamble' we get by trading away 1 apple is:

$$P(\mathrm{heads}) \cdot U(1 \text{ orange}) + P(\mathrm{tails}) \cdot U(3 \text{ plums})$$

$$= 0.50 \cdot 2€ + 0.50 \cdot 1.5€ = 1.75€$$

Conversely, if we just keep our 1 apple instead of making the trade, this has an expected utility of $1 \cdot U(1 \text{ apple}) = €1$. So indeed we ought to trade (as the previous reasoning suggested).

"But wait!" you cry. "Where did these probabilities come from? Why is the 'probability' of a fair coin landing heads 0.5 and not, say, -0.2 or 3? Who says we ought to multiply utilities by probabilities in the first place?"

If you're used to approaching this problem from a Bayesian standpoint, then you may now be thinking of notions like prior probability and Occam's Razor and universal priors...

But from the standpoint of coherence theorems, that's putting the cart before the horse.

From the standpoint of coherence theorems, we don't *start with* a notion of 'probability'.

Instead we ought to prove something along the lines of: if you're not using qualitatively dominated strategies, then you must *behave as if* you are multiplying utilities by certain quantitative thingies.

We might then furthermore show that, for non-dominated strategies, these utility-multiplying thingies must be between 0 and 1 rather than say -0.3 or 27.

Having determined what coherence properties these utility-multiplying thingies need to have, we decide to call them 'probabilities'. And *then*—once we know in the first place that we need 'probabilities' in order to not be using dominated strategies—we can start to worry about exactly what the numbers ought to be.

Probabilities summing to 1

Here's a taste of the kind of reasoning we might do:

Suppose that—having already accepted some previous proof that non-dominated strategies dealing with uncertain outcomes, must multiply utilities by quantitative thingies—you then say that you are going to assign a probability of 0.6 to the coin coming up heads, and a probability of 0.7 to the coin coming up tails.

If you're already used to the standard notion of probability, you might object, "But those probabilities sum to 1.3 when they ought to sum to 1!"[5] But now we are in coherence-land; we don't ask "Did we violate the standard axioms that all the textbooks use?" but "What rules must non-dominated strategies obey?" *De gustibus non est disputandum;* can we *disputandum* somebody saying that a coin has a 60% probability of coming up heads and a 70% probability of coming up tails? (Where these are the only 2 possible outcomes of an uncertain coinflip.)

Well—assuming you've already accepted that we need utility-multiplying thingies—I might then offer you a gamble. How about you give me one apple, and if the coin lands heads, I'll give you 0.8 apples; while if the coin lands tails, I'll give you 0.8 apples.

According to you, the expected utility of this gamble is:

$$P(\text{heads}) \cdot U(0.8 \text{ apples}) + P(\text{tails}) \cdot U(0.8 \text{ apples})$$

$$= 0.6 \cdot 0.8 + 0.7 \cdot 0.8 = 1.04.$$

You've just decided to trade your apple for 0.8 apples, which sure sounds like one of 'em dominated strategies.

And that's why *the thingies you multiply probabilities by*—the thingies that you use to weight uncertain outcomes in your imagination, when you're trying to decide how much you want one branch of an uncertain choice—must sum to 1, whether you call them 'probabilities' or not.

Well... actually we just argued[6] that probabilities for mutually exclusive outcomes should sum to *no more than 1.* What would be an example showing that, for non-dominated strategies, the probabilities for exhaustive outcomes should sum to no less than 1?

Why exhaustive outcomes should sum to at least 1:

Suppose that, in exchange for 1 apple, I credibly offer:

* To pay you 1.1 apples if a coin comes up heads.
* To pay you 1.1 apples if a coin comes up tails.
* To pay you 1.1 apples if anything else happens.

If the probabilities you assign to these three outcomes sum to say 0.9, you will refuse to trade 1 apple for 1.1 apples.

(This is strictly dominated by the strategy of agreeing to trade 1 apple for 1.1 apples.)

Dutch book arguments

Another way we could have presented essentially the same argument as above, is as follows:

Suppose you are a market-maker in a prediction market for some event X. When you say that your price for event X is x, you mean that you will sell for $x a ticket which pays $1 if X happens (and pays out nothing otherwise). In fact, you will sell any number of such tickets!

Since you are a market-maker (that is, you are trying to encourage trading in X for whatever reason), you are also willing to *buy* any number of tickets at the price $x. That is, I can say to you (the market-maker) "I'd like to sign a contract where you give me N · $x now, and in return I must pay you $N iff X happens;" and you'll agree. (We can view this as you selling me a negative number of the original kind of ticket.)

Let X and Y denote two events such that *exactly one* of them must happen; say, X is a coin landing heads and Y is the coin not landing heads.

Now suppose that you, as a market-maker, are motivated to avoid combinations of bets that lead into *certain* losses for you—not just losses that are merely probable, but combinations of bets such that *every* possibility leads to a loss.

Then if exactly one of X and Y must happen, your prices x and y must sum to exactly $1. Because:

- If x + y < $1, I buy both an X-ticket and a Y-ticket and get a guaranteed payout of $1 minus costs of x + y. Since this is a guaranteed profit for me, it is a guaranteed loss for you.
- If x + y > $1, I sell you both tickets and will at the end pay you $1 after you have already paid me x + y. Again, this is a guaranteed profit for me of x + y − $1 > $0.

This is more or less exactly the same argument as in the previous section, with trading apples. Except that: (a) the scenario is more crisp, so it is easier to generalize and scale up much more complicated similar arguments; and (b) it introduces a whole lot of assumptions that people new to expected utility would probably find rather questionable.

"What?" one might cry. "What sort of crazy bookie would buy and sell bets at exactly the same price? Why ought *anyone* to buy and sell bets at exactly the same price? Who says that I must value a gain of $1 exactly the opposite of a loss of $1? Why should the price that I put on a bet represent my degree of uncertainty about the environment? What does all of this argument about gambling have to do with real life?"

So again, the key idea is not that we are assuming anything about people valuing every real-world dollar the same; nor is it in real life a good idea to offer to buy or sell bets at the same prices.[7] Rather, Dutch book arguments can stand in as shorthand for some longer story in which we only assume that you prefer more apples to less apples.

The Dutch book argument above has to be seen as one more added piece in the company of all the *other* coherence theorems—for example, the coherence theorems suggesting that you ought to be quantitatively weighing events in your mind in the first place.

Conditional probability

With more complicated Dutch book arguments, we can derive more complicated ideas such as 'conditional probability'.

Let's say that we're pricing three kinds of gambles over two events Q and R:

- A ticket that costs $x, and pays $1 if Q happens.
- A ticket that doesn't cost anything or pay anything if Q doesn't happen (the ticket price is refunded); and if Q does happen, this ticket costs $y, then pays $1 if R happens.
- A ticket that costs $z, and pays $1 if Q and R both happen.

Intuitively, the idea of conditional probability is that the probability of Q and R both happening, should be equal to the probability of Q happening, times the probability that R happens assuming that Q happens:

$$P(Q \wedge R) = P(Q) \cdot P(R \mid Q)$$

To exhibit a Dutch book argument for this rule, we want to start from the assumption of a qualitatively non-dominated strategy, and derive the quantitative rule $z = x \cdot y$.

So let's give an example that violates this equation and see if there's a way to make a guaranteed profit. Let's say somebody:

- Prices at x = $0.60 the first ticket, aka P(Q).
- Prices at y = $0.70 the second ticket, aka P(R | Q).
- Prices at z = $0.20 the third ticket, aka P(Q ∧ R), which ought to be $0.42 assuming the first two prices.

The first two tickets are priced relatively high, compared to the third ticket which is priced relatively low, suggesting that we ought to sell the first two tickets and buy the third.

Okay, let's ask what happens if we sell 10 of the first ticket, sell 10 of the second ticket, and buy 10 of the third ticket.

- If Q doesn't happen, we get $6, and pay $2. Net +$4.
- If Q happens and R doesn't happen, we get $6, pay $10, get $7, and pay $2. Net +$1.
- If Q happens and R happens, we get $6, pay $10, get $7, pay $10, pay $2, and get $10. Net: +$1.

That is: we can get a guaranteed positive profit over all three possible outcomes.

More generally, let A, B, C be the (potentially negative) amount of each ticket X, Y, Z that is being bought (buying a negative amount is selling). Then the prices x, y, z can be combined into a 'Dutch

book' whenever the following three inequalities can be simultaneously true, with at least one inequality strict:

$$\begin{array}{rrrl} -A\,x & +0 & -C\,z & \geqq 0 \\ A\,(1-x) & -B\,y & -C\,z & \geqq 0 \\ A\,(1-x) & +B\,(1-y) & +C\,(1-z) & \geqq 0 \end{array}$$

For $x, y, z \in (0..1)$ this is impossible exactly iff $z = x \cdot y$. The proof via a bunch of algebra is left as an exercise to the reader.[8]

The Allais Paradox

By now, you'd probably like to see a glimpse of the sort of argument that shows in the first place that we need expected utility—that a non-dominated strategy for uncertain choice must behave as if multiplying utilities by some kinda utility-multiplying thingies ('probabilities').

As far as I understand it, the real argument you're looking for is Abraham Wald's complete class theorem, which I must confess I don't know how to reduce to a simple demonstration.

But we can catch a glimpse of the general idea from a famous psychology experiment that became known as the Allais Paradox (in slightly adapted form).

Suppose you ask some experimental subjects which of these gambles they would rather play:

- 1A: A certainty of $1,000,000.
- 1B: 90% chance of winning $5,000,000, 10% chance of winning nothing.

Most subjects say they'd prefer 1A to 1B.

Now ask a separate group of subjects which of these gambles they'd prefer:

- 2A: 50% chance of winning $1,000,000; 50% chance of winning $0.
- 2B: 45% chance of winning $5,000,000; 55% chance of winning $0.

In this case, most subjects say they'd prefer gamble 2B.

Note that the $ sign here denotes real dollars, not utilities! A gain of five million dollars isn't, and shouldn't be, worth exactly five times as much to you as a gain of one million dollars. We can use the € symbol to denote the expected utilities that are abstracted from how much you relatively value different outcomes; $ is just money.

So we certainly aren't claiming that the first preference is paradoxical because 1B has an expected dollar value of $4.5 million and 1A has an expected dollar value of $1 million. That would be silly. We care about expected utilities, not expected dollar values, and those two concepts aren't the same at all!

Nonetheless, the combined preferences 1A > 1B and 2A < 2B are not compatible with any coherent utility function. We cannot simultaneously have:

$$\begin{array}{rcl} U(\text{gain \$1M}) & > & 0.9 \cdot U(\text{gain \$5M}) + 0.1 \cdot U(\text{gain \$0}) \\ 0.5 \cdot U(\text{gain \$0}) + 0.5 \cdot U(\text{gain \$1M}) & < & 0.45 \cdot U(\text{gain \$5M}) + 0.55 \cdot U(\text{gain \$0}) \end{array}$$

This was one of the earliest experiments seeming to demonstrate that actual human beings were not expected utility maximizers—a very tame idea nowadays, to be sure, but the *first definite*

demonstration of that was a big deal at the time. Hence the term, "Allais Paradox".

Now, by the general idea behind coherence theorems, since we can't *view this behavior* as corresponding to expected utilities, we ought to be able to show that it corresponds to a dominated strategy somehow—derive some way in which this behavior corresponds to shooting off your own foot.

In this case, the relevant idea seems non-obvious enough that it doesn't seem reasonable to demand that you think of it on your own; but if you like, you can pause and try to think of it anyway. Otherwise, just continue reading.

Again, the gambles are as follows:

- 1A: A certainty of $1,000,000.
- 1B: 90% chance of winning $5,000,000, 10% chance of winning nothing.
- 2A: 50% chance of winning $1,000,000; 50% chance of winning $0.
- 2B: 45% chance of winning $5,000,000; 55% chance of winning $0.

Now observe that Scenario 2 corresponds to a 50% chance of playing Scenario 1, and otherwise getting $0.

This, in fact, is why the combination 1A > 1B; 2A < 2B is incompatible with expected utility. In terms of one set of axioms frequently used to describe expected utility, it violates the Independence Axiom: if a gamble L is preferred to M (that is, L > M), then we ought to be able to take a constant probability p > 0 and another gamble N and have

$$p \cdot L + (1 - p) \cdot N > p \cdot M + (1 - p) \cdot N.$$

To put it another way, if I flip a coin to decide whether or not to play some entirely different game N, but otherwise let you choose L or M, you ought to make the same choice as if I just ask you whether you prefer L or M. Your preference between L and M should be 'independent' of the possibility that, instead of doing anything whatsoever with L or M, we will do something else instead.

And since this is an axiom of expected utility, any violation of that axiom ought to correspond to a dominated strategy somehow.

In the case of the Allais Paradox, we do the following:

First, I show you a switch that can be set to A or B, currently set to A.

In one minute, I tell you, I will flip a coin. If the coin comes up heads, you will get nothing. If the coin comes up tails, you will play the gamble from Scenario 1.

From your current perspective, that is, we are playing Scenario 2: since the switch is set to A, you have a 50% chance of getting nothing and a 50% chance of getting $1 million.

I ask you if you'd like to pay a penny to throw the switch from A to B. Since you prefer gamble 2B to 2A, and some quite large amounts of money are at stake, you agree to pay the penny. From your perspective, you now have a 55% chance of ending up with nothing and a 45% chance of getting $5M.

I then flip the coin, and luckily for you, it comes up tails.

From your perspective, you are now in Scenario 1B. Having observed the coin and updated on its state, you now think you have a 90% chance of getting $5 million and a 10% chance of getting nothing. By hypothesis, you would prefer a certainty of $1 million.

So I offer you a chance to pay another penny to flip the switch back from B to A. And with so much money at stake, you agree.

I have taken your two cents on the subject.

That is: You paid a penny to flip a switch and then paid another penny to switch it back, and this is dominated by the strategy of just leaving the switch set to A.

And that's at least a glimpse of why, if you're not using dominated strategies, the thing you do with relative utilities is multiply them by probabilities in a consistent way, and prefer the choice that leads to a greater expectation of the variable representing utility.

From the Allais Paradox to real life

The real-life lesson about what to do when faced with Allais's dilemma might be something like this:

There's *some* amount that $1 million would improve your life compared to $0.

There's some amount that an additional $4 million would further improve your life after the first $1 million.

You ought to visualize these two improvements as best you can, and decide whether another $4 million can produce at least *one-ninth* as much improvement, as much true value to you, as the first $1 million.

If it can, you should consistently prefer 1B > 1A; 2B > 2A. And if not, you should consistently prefer 1A > 1B; 2A > 2B.

The standard 'paradoxical' preferences in Allais's experiment are standardly attributed to a certainty effect: people value the *certainty* of having $1 million, while the difference between a 50% probability and a 55% probability looms less large. (And this ties in to a number of other results about certainty, need for closure, prospect theory, and so on.)

It may sound intuitive, in an Allais-like scenario, to say that you ought to derive some value from being *certain* about the outcome. In fact this is just the reasoning the experiment shows people to be using, so of course it might sound intuitive. But that does, inescapably, correspond to a kind of thinking that produces dominated strategies.

One possible excuse might be that certainty is valuable if you need to make plans about the future; knowing the exact future lets you make better plans. This is admittedly true and a phenomenon within expected utility, though it applies in a smooth way as confidence increases rather than jumping suddenly around 100%. But in the particular dilemma as described here, you only have 1 minute before the game is played, and no time to make other major life choices dependent on the outcome.

Another possible excuse for certainty bias might be to say: "Well, I value the emotional feeling of certainty."

In real life, we do have emotions that are directly about probabilities, and those little flashes of happiness or sadness are worth something if you care about people being happy or sad. If you say that you value the emotional feeling of being *certain* of getting $1 million, the freedom from the fear of getting $0, for the minute that the dilemma lasts and you are experiencing the emotion—well, that may just be a fact about what you value, even if it exists outside the expected utility formalism.

And this genuinely does not fit into the expected utility formalism. In an expected utility agent, probabilities are just thingies-you-multiply-utilities-by. If those thingies start generating their own utilities once represented inside the mind of the person who is an object of ethical value, you really are going to get results that are incompatible with the formal decision theory.

However, *not* being viewable as an expected utility agent does always correspond to employing dominated strategies. You are giving up *something* in exchange, if you pursue that feeling of certainty. You are potentially losing all the real value you could have gained from another $4 million, if that realized future actually would have gained you more than one-ninth the value of the first $1 million. Is a fleeting emotional sense of certainty over 1 minute, worth *automatically* discarding the potential $5-million outcome? Even if the correct answer given your values is that you properly ought to take the $1 million, treasuring 1 minute of emotional gratification doesn't seem like the wise reason to do that. The wise reason would be if the first $1 million really was worth that much more than the next $4 million.

The danger of saying, "Oh, well, I attach a lot of utility to that comfortable feeling of certainty, so my choices are coherent after all" is not that it's mathematically improper to value the emotions we feel while we're deciding. Rather, by saying that the *most valuable* stakes are the emotions you feel during the minute you make the decision, what you're saying is, "I get a huge amount of value by making decisions however humans instinctively make their decisions, and that's much more important than the thing I'm making a decision *about*." This could well be true for something like buying a stuffed animal. If millions of dollars or human lives are at stake, maybe not so much.

Conclusion

The demonstrations we've walked through here aren't the professional-grade coherence theorems as they appear in real math. Those have names like "Cox's Theorem" or "the complete class theorem"; their proofs are difficult; and they say things like "If seeing piece of information A followed by piece of information B leads you into the same epistemic state as seeing piece of information B followed by piece of information A, plus some other assumptions, I can show an isomorphism between those epistemic states and classical probabilities" or "Any decision rule for taking different actions depending on your observations either corresponds to Bayesian updating given some prior, or else is strictly dominated by some Bayesian strategy".

But hopefully you've seen enough concrete demonstrations to get a general idea of what's going on with the actual coherence theorems. We have multiple spotlights all shining on the same core mathematical structure, saying dozens of different variants on, "If you aren't running around in circles or stepping on your own feet or wantonly giving up things you say you want, we can see your behavior as corresponding to this shape. Conversely, if we can't see your behavior as corresponding to this shape, you must be visibly shooting yourself in the foot." Expected utility is the only structure that has this great big family of discovered theorems all saying that. It has a scattering of academic competitors, because academia is academia, but the competitors don't have anything like that mass of spotlights all pointing in the same direction.

So if we need to pick an interim answer for "What kind of quantitative framework should I try to put around my own decision-making, when I'm trying to check if my thoughts make sense?" or "By default and barring special cases, what properties might a sufficiently advanced machine intelligence *look to us* like it possessed, at least approximately, if we couldn't see it *visibly* running around in circles?", then there's pretty much one obvious candidate: Probabilities, utility functions, and expected utility.

Further reading

- To learn more about agents and AI: Consequentialist cognition; the orthogonality of agents' utility functions and capabilities; epistemic and instrumental efficiency; instrumental strategies sufficiently capable agents tend to converge on; properties of sufficiently advanced agents.
- To learn more about decision theory: The controversial counterfactual at the heart of the expected utility formula.

[1] It could be that somebody's pizza preference is real, but so weak that they wouldn't pay one penny to get the pizza they prefer. In this case, imagine we're talking about some stronger preference instead. Like your willingness to pay at least one penny not to have your house burned down, or something.

[2] This does assume that the agent prefers to have more money rather than less money. "Ah, but why is it bad if one person has a penny instead of another?" you ask. If we insist on pinning down every point of this sort, then you can also imagine the $0.01 as standing in for the *time* I burned in order to move the pizza slices around in circles. That time was burned, and nobody else has it now. If I'm an effective agent that goes around pursuing my preferences, I should in general be able to sometimes convert time into other things that I want. In other words, my circular preference can lead me to incur an opportunity cost denominated in the sacrifice of other things I want, and not in a way that benefits anyone else.

[3] There are more than six possibilities if you think it's possible to be absolutely indifferent between two kinds of pizza.

[4] We can omit the 'better doctors' item from consideration: The supply of doctors is mostly constrained by regulatory burdens and medical schools rather than the number of people who want to become doctors; so bidding up salaries for doctors doesn't much increase the total number of doctors; so bidding on a talented doctor at one hospital just means some other hospital doesn't get that talented doctor. It's also illegal to pay for livers, but let's ignore that particular issue with the problem setup or pretend that it all takes place in a more sensible country than the United States or Europe.

[5] Or maybe a tiny bit less than 1, in case the coin lands on its edge or something.

[6] Nothing we're walking through here is really a coherence theorem *per se*, more like intuitive arguments that a coherence theorem ought to exist. Theorems require proofs, and nothing here is what real mathematicians would consider to be a 'proof'.

[7] In real life this leads to a problem of 'adversarial selection', where somebody who knows more about the environment than you can decide whether to buy or sell from you. To put it another way, from a Bayesian standpoint, if an *intelligent* counterparty is deciding whether to buy or sell from you a bet on X, the fact that they choose to buy (or sell) should cause you to update in favor (or against) X actually happening. After all, they wouldn't be taking the bet unless they thought they knew something you didn't!

[8] The quick but advanced argument would be to say that the left-hand-side must look like a singular matrix, whose determinant must therefore be zero.

Integrating disagreeing subagents

In my previous post, I suggested that akrasia involves subagent disagreement - or in other words, different parts of the brain having differing ideas on what the best course of action is. The existence of such conflicts raises the question, how does one resolve them?

In this post I will discuss various techniques which could be interpreted as ways of resolving subagents disagreements, as well as some of the reasons for why this doesn't always happen.

A word on interpreting "subagents"

The frame that I've had so far is that of the brain being composed of different subagents with conflicting beliefs. On the other hand, one could argue that the subagent interpretation isn't strictly necessary for many of the examples that I bring up in this post. One could just as well view my examples as talking about a single agent with conflicting beliefs.

The distinction between these two frames isn't always entirely clear. In "Complex Behavior from Simple (Sub)Agents", mordinamael presents a toy model where an agent has different goals. Moving to different locations will satisfy the different goals to a varying extent. The agent will generate a list of possible moves and picks the move which will bring some goal the closest to being satisfied.

Is this a unified agent, or one made up of several subagents?

One could argue for either interpretation. On the other hand, mordinamael's post frames the goals as subagents, and they are in a sense competing with each other. On the other hand, the subagents arguably don't make the final decision themselves: they just report expected outcomes, and then a central mechanism picks a move based on their reports.

This resembles the neuroscience model I discussed in my last post, where different subsystems in the brain submit various action "bids" to the basal ganglia. Various mechanisms then pick a winning bid based on various criteria - such as how relevant the subsystem's concerns are for the current situation, and how accurate the different subsystems have historically been in their predictions.

Likewise, in extending the model from *Consciousness and the Brain* for my toy version of the Internal Family Systems model, I postulated a system where various subagents vote for different objects to become the content of consciousness. In that model, the winner was determined by a system which adjusted the vote weights of the different subagents based on various factors.

So, subagents, or just an agent with different goals?

Here I would draw an analogy to parliamentary decision-making. In a sense, a parliament as a whole is an agent. Various members of parliament cast their votes, with "the voting system" then "making the final choice" based on the votes that have been cast. That reflects the overall judgment of the parliament as a whole. On the

other hand, for understanding and predicting how the parliament will actually vote in different situations, it is important to model how the individual MPs influence and broker deals with each other.

Likewise, the subagent frame seems most useful when a person's goals interact in such a way that applying the intentional stance - thinking in terms of the beliefs and goals of the individual subagents - is useful for modeling the overall interactions of the subagents.

For example, in my toy Internal Family Systems model, I noted that reinforcement learning subagents might end up forming something like alliances. Suppose that a robot has a choice between making cookies, poking its finger at a hot stove, or daydreaming. It has three subagents: "cook" wants the robot to make cookies, "masochist" wants to poke the robot's finger at the stove, and "safety" wants the robot to *not* poke its finger at the stove.

By default, "safety" is indifferent between "make cookies" and "daydream", and might cast its votes at random. But when it votes for "make cookies", then that tends to avert "poke at stove" more reliably than voting for "daydream" does, as "make cookies" is also being voted for by "cook". Thus its tendency to vote for "make cookies" in this situation gets reinforced.

We can now apply the intentional stance to this situation, and say that "safety" has "formed an alliance" with "cook", as it correctly "believes" that this will avert masochistic actions. If the subagents are also aware of each other and can predict each other's actions, then the intentional stance gets even more useful.

Of course, we could just as well apply the purely mechanistic explanation and end up with the same predictions. But the intentional explanation often seems easier for humans to reason with, and helps highlight salient considerations.

Integrating beliefs, naturally or with techniques

In any case, regardless of whether we are talking about subagents with conflicting beliefs or just conflicting goals, it still seems like many of our problems arise from *some* kind of internal disagreement. I will use the term "integration" for anything that acts to resolve such conflicts, and discuss a few examples of things which can be usefully thought of as integration.

In these examples, I am again going to rely on the basic observation from *Consciousness and the Brain*: that when some subsystem in the brain manages to elevate a mental object into the content of consciousness, multiple subsystems will synchronize their processing around that object. Assuming that the conditions are right, this will allow for the integration of otherwise conflicting beliefs or behaviors.

Why do we need to explicitly integrate beliefs, rather than this happening automatically? One answer is that trying to integrate all beliefs would be infeasible; as CronoDAS notes:

> GEB has a section on this.

> In order to *not* compartmentalize, you need to test if your beliefs are all consistent with each other. If your beliefs are all statements in propositional logic, consistency checking becomes the Boolean Satisfiability Problem, which is NP-complete. If your beliefs are statements in predicate logic, then consistency checking becomes PSPACE-complete, which is even worse than NP-complete.

Rather than try to constantly integrate every possible belief and behavior, the brain will rather try to integrate beliefs at times when it notices contradictions. Of course, sometimes we *do* realize that there are contradictions, but still don't automatically integrate the subagents. Then we can use various techniques for making integration more effective. How come integration isn't more automatic?

One reason is that integration requires the right conditions, and while the brain has mechanisms for getting those conditions right, integration is still a nontrivial skill. As an analogy, most children learn the basics of talking and running on their own, but they can still explicitly study rhetoric or running techniques to boost their native competencies far above their starting level. Likewise, everyone natively does *some* integration on their own, but people can also use explicit techniques which make them much better at it.

Resisting belief integration

Lack of skill isn't the full answer for why we don't always automatically update, however. Sometimes it seems as if the mind actively resists updating.

One of the issues that commonly comes up in Internal Family Systems therapy is that parts of the mind want to keep some old belief frozen, because if it were known, it would change the person's behavior in an undesired way. For example, if someone believes that they have a good reason not to abandon their friend, then a part of the mind which values not abandoning the friend in question might resist having this belief re-revaluated. The part may then need to be convinced that knowing the truth only leaves opens the *option* of abandoning the friend, it doesn't *compel* it.

Note that this isn't *necessarily* true. If there are other subagents which sufficiently strongly hold the opinion that the friend should be abandoned, and the subagent-which-values-the-friend is only managing to prevent that by hanging on to a specific belief, then readjusting that belief *might* remove the only constraint which was preventing the anti-friend coalition from dumping the friend. Thus from the point of view of the subagent which is resisting the belief update, the update *would* compel an abandonment of the friend. In such a situation, additional internal work may be necessary before the subagent will agree to let the belief revision proceed.

More generally, subagents may be incentivized to resist belief updating for at least three different reasons (this list is not intended to be exhaustive):

1. The subagent is trying to pursue or maintain a goal, and predicts that revising some particular belief would make the person less motivated to pursue or maintain the goal.
2. The subagent is trying to safeguard the person's social standing, and predicts that not understanding or integrating something will be safer, give the person an advantage in negotiation, or be otherwise socially beneficial. For instance, different subagents holding conflicting beliefs allows a person to verbally believe

in one thing while still not acting accordingly - even actively changing their verbal model so as to avoid falsifying the invisible dragon in the garage.
3. Evaluating a belief would require activating a memory of a traumatic event that the belief is related to, and the subagent is trying to keep that memory suppressed as part of an exile-protector dynamic.

Here's an alternate way of looking at the issue, which doesn't use the subagent frame. So far I have been mostly talking about integrating beliefs rather than goals, but humans don't seem to have a clear value/belief distinction. As Stuart Armstrong discusses in his mAIry's room article, for humans simply receiving sensory information often also rewires some of their values. Now, Mark Lippman suggests that trying to optimize a complicated network of beliefs and goals means that furthering one goal may hurt other goals, so the system needs to have checks in place to ensure that one goal is not pursued in a way which disproportionately harms the achievement of other goals.

For example, most people wouldn't want to spend the rest of their lives doing nothing but shooting up heroin, *even if* they knew for certain that this maximized the achievement of their "experience pleasure" goal. If someone offered them the chance to experience just *how* pleasurable heroin felt like - giving them more accurate emotion-level predictions of the experience - they might quite reasonably refuse, as they feared that making this update might make them more inclined to take heroin. Eliezer once noted that if someone offered him a pill which simulated the joy of scientific discovery, he would make sure never to take it.

Suppose that a system has a network of beliefs and goals and it does something like predicting how various actions and their effects - not only their effects on the external world, but on the belief/goal network itself - might influence its goal achievement. If it resists actions which reduce the probability of achieving its current goals, then this might produce dynamics which look like subagents trying to achieve their goals at the expense of the other subagents.

For instance, Eliezer's refusal to take the pill might be framed as a subagent valuing scientific discovery trying to block a subagent valuing happiness from implementing an action which would make the happiness subagent's bids for motor system access stronger. Alternatively, it might be framed as the overall system putting value on actually making scientific discoveries, and refusing to self-modify in a way which it predicted would hurt this goal. (You might note that this has some interesting similarities to things like the Cake or Death problem in AI alignment.)

In any case, integration is not always straightforward. Even if the system *does* detect a conflict between its subagents, it may have a reason to avoid doing so.

Having reviewed some potential barriers for integration, let us move on to different ways in which conflicts *can* be detected and integrated.

Ways to integrate conflicting subagents

Cognitive Behavioral Therapy

Scott Alexander has an old post where he quotes this excerpt from the cognitive behavioral therapy book *When Panic Attacks*:

> I asked Walter how he was thinking and feeling about the breakup with Paul. What was he telling himself? He said "I feel incredibly guilty and ashamed, and it seems like it must have been my fault. Maybe I wasn't skillful enough, attractive enough, or dynamic enough. Maybe I wasn't there for him emotionally. I feel like I must have screwed up. Sometimes I feel like a total fraud. Here I am, a marriage and family therapist, and my own relationship didn't even work out. I feel like a loser. A really, really big loser." [...]
>
> I thought the Double Standard Technique might help because Walter seemed to be a warm and compassionate individual. I asked what he'd say to a dear friend who'd been rejected by someone he'd been living with for eight years. I said "Would you tell him that there's something wrong with him, that he screwed up his life and flushed it down the toilet for good?"
>
> Walter looked shocked and said he'd never say something like that to a friend. I suggested we try a role-playing exercise so that he could tell me what he would say to a friend who was in the same predicament [...]
>
> **Therapist (role-playing patient's friend):** Walter, there's another angle I haven't told you about. What you don't understand is that I'm impossible to live with and be in a relationship with. That's the real reason I feel so bad, and that's why I'll be alone for the rest of my life.
>
> **Patient (role-playing as if therapist is his friend who just had a bad breakup):** Gosh, I'm surprised to hear you say that, because I've known you for a long time and never felt that way about you. In fact, you've always been warm and open, and a loyal friend. How in the world did you come to the conclusion that you were impossible to be in a relationship with?
>
> **Therapist (continuing role-play):** Well, my relationship with [my boyfriend] fell apart. Doesn't that prove I'm impossible to be in a relationship with?
>
> **Patient (continuing role-play):** In all honesty, what you're saying doesn't make a lot of sense. In the first place, your boyfriend was also involved in the relationship. It takes two to tango. And in the second place, you were involved in a reasonably successful relationship with him for eight years. So how can you claim that you're impossible to live with?
>
> **Therapist (continuing role-play:)** Let me make sure I've got this right. You're saying that I was in a reasonably successful relationship for eight years, so it doesn't make much sense to say that I'm impossible to live with or impossible to be in a relationship with?
>
> **Patient (continuing-role-play:)** You've got it. Crystal clear.
>
> At that point, Walter's face lit up, as if a lightbulb had suddenly turned on in his brain, and we both started laughing. His negative thoughts suddenly seemed absurd to him, and there was an immediate shift in his mood…after Walter put the lie to his negative thoughts, I asked him to rate how he was feeling again. His feeling of sadness fell all the way from 80% to 20%. His feelings of guilt, shame, and anxiety fell all the way to 10%, and his feelings of hopelessness dropped to

> 5%. The feelings of loneliness, embarrassment, frustration, and anger disappeared completely.

At the time, Scott expressed confusion about how just telling someone that their beliefs aren't rational, would be enough to transform the beliefs. But that wasn't really what happened. Walter was asked whether he'd say something harsh to a friend, and he said no, but that alone wasn't enough to improve his condition. What did help was putting him in a position where he had to really think through the arguments for why this is irrational in order to convince his friend, and then, after having formulated the arguments once himself, get convinced by them himself.

In terms of our framework, we might say that a part of Walter's mind contained a model which output a harsh judgment of himself, while another part contained a model which would output a much less harsher judgment of someone else who was in otherwise identical circumstances. Just bringing up the existence of this contradiction wasn't enough to change it: it caused the contradiction to be *noticed*, but didn't activate the relevant models extensively enough for their contents to be reprocessed.

But when Walter had to role-play a situation where he thought of himself as *actually* talking with a depressed friend, that required him to more fully activate the non-judgmental model and apply it to the relevant situation. This caused him to blend with the model, taking its perspective as the truth. When that perspective was then propagated to the self-critical model, the easiest way for the mind to resolve the conflict was simply to alter the model producing the self-critical thoughts.

Note that this kind of a result wasn't *guaranteed* to happen: Walter's self-critical model might have had a reason for why these cases were actually different, and pointing out that reason would have been another way for the contradiction to be resolved. In the example case, however, it seemed to work.

Mental contrasting

Another example of activating two conflicting mental models and forcing an update that way comes from the psychologist Gabriele Oettingen's book Rethinking Positive Thinking. Oettingen is a psychologist who has studied combining a mental imagery technique known as "mental contrasting" with trigger-action planning.

It is worth noting that this book has come under some heavy criticism and may be based on cherry-picked studies. However, in the book this particular example is just presented as an anecdote without even trying to cite any particular studies in its support. I present it because I've personally found the technique to be useful, and because it feels like a nice concise explanation of the kind of integration that often works:

> Try this exercise for yourself. Think about a fear you have about the future that is vexing you quite a bit and that you know is unjustified. Summarize your fear in three to four words. For instance, suppose you're a father who has gotten divorced and you share custody with your ex-wife, who has gotten remarried. For the sake of your daughter's happiness, you want to become friendly with her stepfather, but you find yourself stymied by your own emotions. Your fear might be "My daughter will become less attached to me and more attached to her stepfather." Now go on to imagine the worst possible outcome. In this case, it might be "I feel distanced from my daughter. When I see her she ignores me, but she eagerly

> spends time with her stepfather." Okay, now think of the positive reality that stands in the way of this fear coming true. What in your actual life suggests that your fear won't really come to pass? What's the single key element? In this case, it might be "The fact that my daughter is extremely attached to me and loves me, and it's obvious to anyone around us." Close your eyes and elaborate on this reality.
>
> Now take a step back. Did the exercise help? I think you'll find that by being reminded of the positive reality standing in the way, you will be less transfixed by the anxious fantasy. When I conducted this kind of mental contrasting with people in Germany, they reported that the experience was soothing, akin to taking a warm bath or getting a massage. "It just made me feel so much calmer and more secure," one woman told me. "I sense that I am more grounded and focused."
>
> Mental contrasting can produce results with both unjustified fears as well as overblown fears rooted in a kernel of truth. If as a child you suffered through a couple of painful visits to the dentist, you might today fear going to get a filling replaced, and this fear might become so terrorizing that you put off taking care of your dental needs until you just cannot avoid it. Mental contrasting will help you in this case to approach the task of going to the dentist. But if your fear is justified, then mental contrasting will confirm this, since there is nothing preventing your fear from coming true. The exercise will then help you to take preventive measures or avoid the impending danger altogether.

As in the CBT example, first one mental model (the one predicting losing the daughter's love) is activated and intentionally blended with, after which an opposing one is, forcing integration. And as in Walter's example, this is not *guaranteed* to resolve the conflict in a more reassuring way: the mind can also resolve the conflict by determining that actually the fear *is* justified.

Internal Double Crux / Internal Family Systems

On some occasions a single round of mental contrasting, or the Walter CBT technique, might be enough. In that case, there were two disagreeing models, and bringing the disagreement into consciousness was enough to reject the other one entirely. But it is not always so clear-cut; sometimes there are subagents which disagree, and both of them actually have some valid points.

For instance, someone might have a subagent which wants the person to do socially risky things, and another subagent which wants to play things safe. Neither is unambiguously wrong: on the other hand, some things *are* so risky that you should never try to do them. On the other hand, *never* doing anything which others might disapprove of is not going to lead to a particularly happy life, either.

In that case, one may need to actively facilitate a *dialogue* between the subagents, such as in the CFAR technique of Internal Double Crux (description, discussion and example, example as applied to dieting), iterating it for several rounds until both subagents come to agreement. The CBT and mental contrasting examples above might be considered special cases of an IDC session, where agreement was reached within a single round of discussion.

More broadly, IDC itself can be considered a special case of applying Internal Family Systems, which includes facilitating conversations between mutually opposing subagents as one of its techniques.

Self-concept editing

In the summer of 2017, I found Steve Andreas's book *Transforming Your Self*, and applied its techniques to fixing a number of issues in my self-concepts which had contributed to my depression and anxiety. Effects from this work which have lasted include no longer having generalized feelings of shame, no longer needing constant validation to avoid such feelings of shame, no longer being motivated by a desire to prove to myself that I'm a good person, and no longer having obsessive escapist fantasies, among other things.

I wrote an article at the time that described the work. The model in *Transforming Your Self* is that I might have a self-concept such as "I am kind". That self-concept is made up of memories of times when I either was kind (*examples* of the concept), or times when I was not (*counterexamples*). In a healthy self-concept, both examples and counterexamples are integrated together: you might have memories of how you are kind in general, but also memories of not being very kind at times when you were e.g. under a lot of stress. This allows you to both know your general tendency, as well as letting you prepare for situations where you know that you won't be very kind.

The book's model also holds that sometimes a person's counterexamples might be split off from their examples. This leads to an unstable self-concept: either your subconscious attention is focused on the examples and totally ignores the counterexamples, in which case you feel good and kind, or it swings to the counterexamples and totally ignores the examples, in which case you feel like a terrible horrible person with no redeeming qualities. You need a constant stream of external validation and evidence in order to keep your attention anchored on the examples; the moment it ceases, your attention risks swinging to the counterexamples again.

While I didn't have the concept back then, what I did could also be seen as integrating true but disagreeing perspectives between two subagents. There was one subagent which held memories of times when I had acted in what it thought of as a bad way, and was using feelings of shame to motivate me to make up for those actions. Another subagent was then reacting to it by making me do more and more things which I could use to prove to myself and others that I was indeed a good person. (This description roughly follows the framing and conceptualization of self-esteem and guilt/shame in the IFS book *Freedom from your Inner Critic*.)

Under the sociometer theory of self-esteem, self-esteem is an internal evaluation of one's worth as a partner to others. With this kind of an interpretation, it makes sense to have subagents acting in the ways that I described: if you have done things that your social group would judge you for, then it becomes important to do things which prove your worth and make them forgive you.

This then becomes a special case of an IFS exile/protector dynamic. Under that formulation, the splitting of the counterexamples and the lack of updating actually serves a purpose. The subagent holding the memories of doing shameful things doesn't want to stop generating the feelings of shame until it has received sufficient evidence that the "prove your worth" behavior has actually become unnecessary.

One of the techniques from *Transforming Your Self* that I used to fix my self-concept was integrating the examples by adding qualifiers to the counterexamples: “when I was a child, and my executive control wasn’t as developed, I didn’t always act as kindly as I could have”. Under the belief framing, this allowed my memories to be integrated in a way which showed that my selfishness as a child was no longer evidence of me being horrible in general. Under the subagent framing, this communicated to the shame-generating subagent that the things that I did as a child would no longer be held against me, and that it was safe to relax.

Another technique mentioned in *Transforming Your Self*, which I did not personally need to use, was translating the concerns of subagents into a common language. For instance, someone’s positive self-concept examples might be in the form of mental images, with their negative counterexamples being in the form of a voice which reminds them of their failures. In that case, they might translate the inner speech into mental imagery by visualizing what the voice is saying, turning both the examples and counterexamples into mental images that can then be combined. This brings us to…

Translating into a common language

Eliezer presents an example of two different framings eliciting conflicting behavior in his “Circular Altruism” post:

> Suppose that a disease, or a monster, or a war, or something, is killing people. And suppose you only have enough resources to implement one of the following two options:
>
> 1. Save 400 lives, with certainty.
>
> 2. Save 500 lives, with 90% probability; save no lives, 10% probability.
>
> Most people choose option 1. […] If you present the options this way:
>
> 1. 100 people die, with certainty.
>
> 2. 90% chance no one dies; 10% chance 500 people die.
>
> Then a majority choose option 2. *Even though it's the same gamble*. You see, just as a *certainty* of saving 400 lives seems to *feel* so much more comfortable than an unsure gain, so too, a certain loss *feels* worse than an uncertain one.

In my previous post, I presented a model where subagents which are most strongly activated by the situation are the ones that get access to the motor system. If you are hungry and have a meal in front of you, the possibility of eating is the most salient and valuable feature of the situation. As a result, subagents which want you to eat get the most decision-making power. On the other hand, if this is a restaurant in Jurassic Park and a velociraptor suddenly charges through the window, then the dangerous aspects of the situation become most salient. That lets the subagents which want you to flee to get the most decision-making power.

Eliezer’s explanation of the saving lives dilemma is that in the first framing, the certainty of saving 400 lives is salient, whereas in the second explanation the certainty of losing 100 lives is salient. We can interpret this in similar terms as the “eat or run” dilemma: the action which gets chosen, depends on which features are

the most salient and how those features activate different subagents (or how those features highlight different priorities, if we are not using the subagent frame).

Suppose that you are someone who was tempted to choose option 1 when you were presented with the first framing, and option 2 when you were presented with the second framing. It is now pointed out to you that these are actually exactly equivalent. You realize that it would be inconsistent to prefer one option over the other just depending on the framing. Furthermore, and maybe even more crucially, realizing this makes *both* the "certainty of saving 400 lives" *and* "certainty of losing 100 lives" features become equally salient. That puts the relevant subagents (priorities) on more equal terms, as they are both activated to the same extent.

What happens next depends on what the relative strengths of those subagents (priorities) are otherwise, and whether you happen to know about expected value. Maybe you consider the situation and one of the two subagents (priorities) happens to be stronger, so you decide to consistently save 400 or consistently lose 100 lives in both situations. Alternatively, the conflicting priorities may be resolved by introducing the rule that "when detecting this kind of a dilemma, convert both options into an expected value of lives saved, and pick the option with the higher value".

By converting the options to an expected value, one can get a basis by which two otherwise equal options can be evaluated and chosen between. Another way of looking at it is that this is bringing in a third kind of consideration/subagent (knowledge of the decision-theoretically optimal decision) in order to resolve the tie.

Urge propagation

CFAR and Harvard Effective Altruism is a video of a lecture given by former CFAR instructors Valentine Smith and Duncan Sabien. In Valentine's part of the lecture, he describes a few motivational techniques which work by mentally reframing the contents of an experience.

The first example involves having a $50 parking ticket, which - unless paid within 30 days - will accrue an additional $90 penalty. This kind of a thing tends to feel ughy to deal with, causing an inclination to avoid thinking about it - while also being aware of the need to do something about it. Something along the lines of two different subagents which are both trying to avoid pain using opposite methods - one by not thinking about unpleasant things, another by doing things which stop future unpleasantness.

Val's suggested approach involves noting that if you instead had a cheque for $90, which would expire in 30 days, then that would *not* cause such a disinclination. Rather, it would feel actively pleasant to cash it in and get the money.

The structure of the "parking ticket" and "cheque" scenarios are equivalent, in that both cases you can take an action to be $90 better off after 30 days. If you notice this, then it may be possible for you to re-interpret the action of paying off the parking ticket as something that *gains you money*, maybe by something like literally looking at it and imagining it as a cheque that you can cash in, until cashing it in starts feeling *actively pleasant*.

Val emphasizes that this is not just an arbitrary motivational hack: it's important that your reframe is *actually bringing in real facts from the world*. You don't want to just

imagine the parking ticket as a ticking time bomb, or as something else which it actually isn't. Rather, you want to do a reframe which integrates both perspectives, while also highlighting the features which will help fix the conflict.

One description of what happens here would be that once the pain-avoiding subagent notices that paying the parking ticket can feel like a net gain, and that it being a net gain is actually describing a real fact about the world, then it can drop its objection and you can proceed to take actions. The other way of looking at it is that like with expected value, you are introducing a common currency - the future impact on your finances - which allows the salient features from both subagents' perspectives to be integrated and then resolved.

Val's second example involves a case where he found himself not doing push-ups like he had intended to. When examining the reason why not, he noticed that the push-ups felt physically unpleasant: they involved sweating, panting, and a burning sensation, and this caused a feeling of aversion.

Part of how he solved the issue was by realizing that his original goal for getting exercise was to live longer and be in better health. The unpleasant physical sensations were a sign that *he was pushing his body hard enough that the push-ups would actually be useful for this goal*. He could then create a mental connection between the sensations and his goal of being healthier and living longer: the sensations started feeling like *something positive*, since they were an indication of progress.

Besides being an example of creating a common representation between the subagents, this can also be viewed as doing a round of Internal Double Crux, something like:

> **Exercise subagent:** We should exercise.
>
> **Optimizer subagent:** That feels unpleasant and costs a lot of energy, we would have the energy to do more things if we didn't exercise.
>
> **Exercise subagent:** That's true. But the feelings of unpleasantness are actually a sign of us getting more energy in the long term.
>
> **Optimizer subagent:** Oh, you're right! Then let's exercise, that furthers my goals too.

(There's also a bunch of other good stuff in the video that I didn't describe here, you may want to check it out if you haven't already done so.)

Exposure Therapy

So far, most of the examples have assumed that the person already has all the information necessary for solving the internal disagreement. But sometimes additional information might be required.

The prototypical use of *exposure therapy* is for phobias. Someone might have a phobia of dogs, while at the same time feeling that their fear is irrational, so they decide to get therapy for their phobia.

How the therapy typically proceeds is by exposing the person to their fear in increments that are as small as possible. For instance, a page by Anxiety Canada

offers this list of steps that someone might have for exposing themselves to dogs:

Step 1: Draw a dog on a piece of paper.

Step 2: Read about dogs.

Step 3: Look at photos of dogs.

Step 4: Look at videos of dogs.

Step 5: Look at dogs through a closed window.

Step 6: Then through a partly-opened window, then open it more and more.

Step 7: Look at them from a doorway.

Step 8: Move further out the doorway; then further etc.

Step 9: Have a helper bring a dog into a nearby room (on a leash).

Step 10: Have the helper bring the dog into the same room, still on a leash.

The ideal is that each step is enough to make you feel a little scared, but not so scared that it would serve to act retraumatize you or otherwise make you feel horrible about what happened.

In a sense, exposure therapy involves one part of the mind thinking that the situation is safe, and another part of the mind thinking that the situation is unsafe, and the contradiction being resolved by *testing* it. If someone feels nervous about looking at a photo of a dog, it implies that a part of their mind thinks that seeing a photo of a dog means they are potentially in danger. (In terms of the machine learning toy model from my IFS post, it means that a fear model is activated, which predicts the current state to be dangerous.)

By looking at photos sufficiently many times, and then afterwards noting that everything is okay, the nervous subagent gets information about having been wrong, and updates its model. Over time, and as the person goes forward in steps, the nervous subagent can eventually conclude that it had overgeneralized from the original trauma, and that dogs in general aren't that dangerous after all.

As in the CBT example, one can view this as activating conflicting models and the mind then fixing the conflict by updating the models. In this case, the conflict is between the frightened subagent's prediction that seeing the dog is a sign of danger, and another subagent's later assessment that everything turned out to be fine.

Conclusion to integration methods

I have considered here a number of ways of integrating subagent conflicts. Here are a few key principles that are used in them:

- **Selectively blending with subagents/beliefs to make disagreements between them more apparent.** Used in the Cognitive Behavioral Therapy and mental contrasting cases. Also used in a somewhat different form in exposure therapy, where you are partially blended with a subagent that thinks that the

situation is dangerous, while getting disagreeing information from the rest of the world.
- **Facilitating a dialogue between subagents “from the outside”.** Used in Internal Double Crux, Internal Family Systems. In a sense, the next bullet can also be viewed a special case of this.
    - **Combining aspects of the conflicting perspectives to a whole which allows for resolution.** Used in self-concept editing, Eliezer’s altruism example, and urge propagation.
- **Collecting additional information which allows for the disagreement to be resolved.** Used in exposure therapy.

I believe that we have evolved to use all of these spontaneously, without necessarily realizing what it is that we are doing.

For example, many people have the experience of it being useful to talk to a friend about your problems, weighting the pros and cons of different options. Frequently just getting to talk about it helps clarify the issue, even if the friend doesn’t say anything (or even if they are a rubber duck). Probably not coincidentally, if you are talking about the conflicting feelings that you have in your mind, then you are frequently doing something like an informal version of Internal Double Crux. You are representing all the sides of a dilemma until you have reached a conclusion and integrated the different perspectives.

To the extent that they are effective, various schools of therapy and self-improvement - ranging from CBT to IDC to IFS - are formalized methods for making such integration more effectively.

Separation of Concerns

Separation of concerns is a principle in computer science which says that distinct concerns should be addressed by distinct subsystems, so that you can optimize for them separately. We can also apply the idea in many other places, including human rationality. This idea has been written about before. I'm not trying to make a comprehensive post about it, just remark on some things I recently though about.

Epistemic vs Instrumental

The most obvious example is beliefs vs desires. Although the distinction may not be a *perfect* separation-of-concerns in practice (or even in principle), at least I can say this:

- Even non-rationalists find it useful to make a relatively firm distinction between what is true and what they want to be true;
- Rationalists, scientists, and intellectuals of many varieties tend to value an especially sharp distinction of this kind.

I'm particularly thinking about how the distinction is used in conversation. If an especially sharp distinction *isn't* being made, you might see things like:

- Alice makes a factual statement, but the statement has (intended or unintended) conversational implicature which is perceived as negative by most of the people present. Alice is chastised and concedes the point, withdrawing her assertion.
- Bob mentions a negative consequence of a proposed law. Everyone listening perceives Bob to be arguing against the law.

Notice that this isn't an easy distinction to make. It isn't right at all to just ignore conversational implicature. You should not only make literal statements, nor should you just assume that everyone else is doing that. The skill is more like, raise the literal content of words *as a hypothesis*; make a distinction in your mind between what is said and anything else which may have been meant.

Side note -- as with many conversation norms, the distinctions I'm mentioning in this post cannot be imposed on a conversation unilaterally. Sometimes simply pointing out a distinction works; but generally, one has to meet a conversation where it's at, and only gently try to pull it to a better place. If you're in a discussion which is strongly failing to make a true-vs-useful distinction, simply pointing out examples of the problem will very likely be taken as an attack, making the problem worse.

Making a distinction between epistemics and instrumentality seems like a kind of "universal solvent" for cognitive separation of concerns -- the rest of the examples I'm going to mention feel like consequences of this one, to some extent. I think part of the reason for this is that "truth" is a concept which has a lot of separation-of-concerns built in: it's not *just* that you consider truth separately from usefulness; you also consider the truth of *each individual statement* separately, which creates a scaffolding to support a huge variety of separation-of-concerns (any time you're able to make an explicit distinction between different assertions).

But the distinction is also very broad. Actually, it's kind of a mess -- it feels a bit like "truth vs everything else". Earlier, I tried to characterize it as "what's true vs what you want to be true", but taken literally, this only captures a narrow case of what I'm pointing at. There are many different goals which statements can optimize besides truth.

- You could want to believe something because you want it to be true -- perhaps you can't stand thinking about the possibility of it being false.
- You could want to claim something because it helps argue for/against some side in a decision which you want to influence, or for/against some other belief which you want to hold for some other reason.
- You could want to believe something because the behaviors encouraged by the belief are good -- perhaps you exercise more if you believe it will make you lose weight; perhaps everyone believing in karma, or heaven and hell, makes for a stronger and more cooperative community.

Simply put, there are a wide variety of incentives on beliefs and claims. There wouldn't even be a concept of 'belief' or 'claim' if we didn't separate out the idea of truth from all the other reasons one might believe/claim something, and optimize for it separately. Yet, it is kind of fascinating that we do this even to the degree that we do -- how do we successfully identify the 'truth' concern in the first place, and sort it out from all the other incentives on our beliefs?

Argument vs Premises and Conclusion

Another important distinction is to separate the evaluation of hypothetical if-then statements from any concern with the truth of their premises or conclusions. A common complaint among the more logic-minded, of the less, is that hardly anyone is capable of properly distinguishing the claim "If X, then Y" from the claim "X, and also Y".

It could be that a lack of a very sharp truth-vs-implicature distinction is what blocks people from making an if-vs-and distinction. Why would you be claiming "If X, then Y" if not to then say "by the way, X; so, Y"? (There are actually lots of reasons, but, they're all much less common than making an argument because you believe the premises and want to argue the conclusion -- so, that's the commonly understood implicature.)

However, it's also possible to successfully make the "truth" distinction but not the "hypothetical" distinction. Hypothetical reasoning is a tricky skill. Even if you successfully make the distinction when it is pointed out explicitly, I'd guess that there are times when you fail to make it in conversation or private thought.

Preferences vs Bids

The main reason I'm writing this post is actually because this distinction hit me recently. You can say that you want something, or say how you feel about something, without it being a bid for someone to do something about it. This is both close to the overall topic of In My Culture and a specific example (like, listed as an example in the post).

Actually, let's split this up into cases:

Preferences about social norms vs bids for those social norms to be in place. This is more or less the point of the In My Culture article; saying "in my culture" before something to put a little distance between the conversation and the preferred norm, so that it is put on the table as an invitation rather than being perceived as a requirement.

Proposals vs preferences vs bids. Imagine a conversation about what restaurant to go to. Often, people run into a problem: no one has any preferences; everyone is fine with whatever. No one is willing to make any proposals. One reason why this might happen is that proposals, and preferences, are perceived as bids. No one wants to take the blame for a bad plan; no one wants to be seen as selfish or negligent of other's preferences. So, there's a natural inclination to lose touch with your preferences; you *really feel* like you don't care, and like you can't think of any options.

- A *proposal* puts an option 'on the table' for consideration.
- A *preference* is your own component of the group utility function. If you also think other people should have the same preference, you can state your reason for that, and let others update if they will.
- A *bid* is a request for group action: you don't just *want* tacos, you don't even merely *propose* tacos; you *call on the group to collectively get tacos.*

If a strong distinction between preferences and bids is made, it gets easier to state what you prefer, trusting that the group will take it as only one data point of many to be taken together. If a distinction between proposals and bids is made, it will be easier to list whatever comes to mind, and to think of places you'd actually like to go.

Feelings vs bids. I think this one comes less naturally to people who make a strong truth distinction -- there's something about directing attention toward the literal truth of statements which directs attention away from how you feel about them, even though how you feel is something you can also try to have true beliefs about. So, in practice, people who make an especially strong truth distinction may nonetheless treat statements about feelings as if they were statements about the things the feelings are about, precisely because they're hypersensitive to other people failing to make that distinction. So: *know that you can say how you feel about something without it being anything more.* Feeling angry about someone's statement doesn't have to be a bid for them to take it back, or a claim that it is false. Feeling sad doesn't have to be a bid for attention. An emotion doesn't even have to reflect your more considered preferences.

(To make this a reality, you probably have to explicitly flag that your emotions are *not* bids.)

When a group of people is skilled at making a truth distinction, certain kinds of conversation, and certain kinds of thinking, become much easier: all sorts of beliefs can be put out into the open where they otherwise couldn't, allowing the collective knowledge to go much further. Similarly, when a group of people is skilled at the *feelings* distinction, I expect things can go places where they otherwise couldn't. If you can mention in passing that something everyone else seems to like makes you sad, without it becoming a big deal. If there is sufficient trust that you can say how you are feeling about things, *in detail*, without expecting it to make everything complicated.

The main reason I wrote this post is that someone was talking about this kind of interaction, and I initially didn't see it as very possible or necessarily desirable. After thinking about it more, the analogy to making a strong truth distinction hit me.

Someone stuck in a culture without a strong truth distinction might similarly see such a distinction as 'not possible or desirable': the usefulness of an assertion is obviously more important than its truth; in reality, being overly obsessed with truth will both make you vulnerable (if you say true things naively) and ignorant (if you take statements at face value too much, ignoring connotation and implicature); even if it were possible to set aside those issues, what's the use of saying a bunch of true stuff? Does it get things done? Similarly: the truth of the matter is more important than how you feel about it; in reality, stating your true feelings all the time will make you vulnerable and perceived as needy or emotional; even if you could set those things aside, what's the point of talking about feelings all the time?

Now it seems both are possible and simply good, for roughly the same reason. Having the ability to make distinctions doesn't require you to explicitly point out those distinctions in every circumstance; rather, it opens up more possibilities.

I can't say a whole lot about the benefits of a feelings-fluent culture, because I haven't really experienced it. This kind of thing is part of what circling seems to be about, in my mind. I think the rationalist community as I've experienced it goes *somewhat* in that direction, but definitely not all the way.

Naked mole-rats: A case study in biological weirdness

This is a linkpost for https://eukaryotewritesblog.com/2019/05/19/naked-mole-rats-a-case-study-in-biological-weirdness/

Epistemic status: Speculative, just having fun. This piece isn't well-cited, but I can pull up sources as needed - nothing about mole-rats is my original research. A lot of this piece is based on Wikipedia.

When I wrote about "weirdness" in the past, I called marine invertebrates, archaea viruses, and Florida Man stories "predictably weird". This means I wasn't really surprised to learn any new wild fact about them. But there's a sense in which marine invertebrates both are and aren't weird. I want to try operationalizing "weirdness" as "amount of unpredictability or diversity present in a class" (or "in an individual") compared to other members of its group.

So in terms of "animals your hear about" - well, you know the tigers, the mice, the bees, the tuna fish, the songbirds, whatever else comes up in your life. But "deep sea invertebrates" seems to include a variety of improbable creatures - a betentacled neon sphere covered in spikes, a six-foot long disconcertingly smooth and flesh-colored worm, bisexual squids, etc. Hey! Weird! That's weird.

But looking at a phylogenetic tree, we see really quickly that "invertebrates" represent almost the entire animal tree of life.

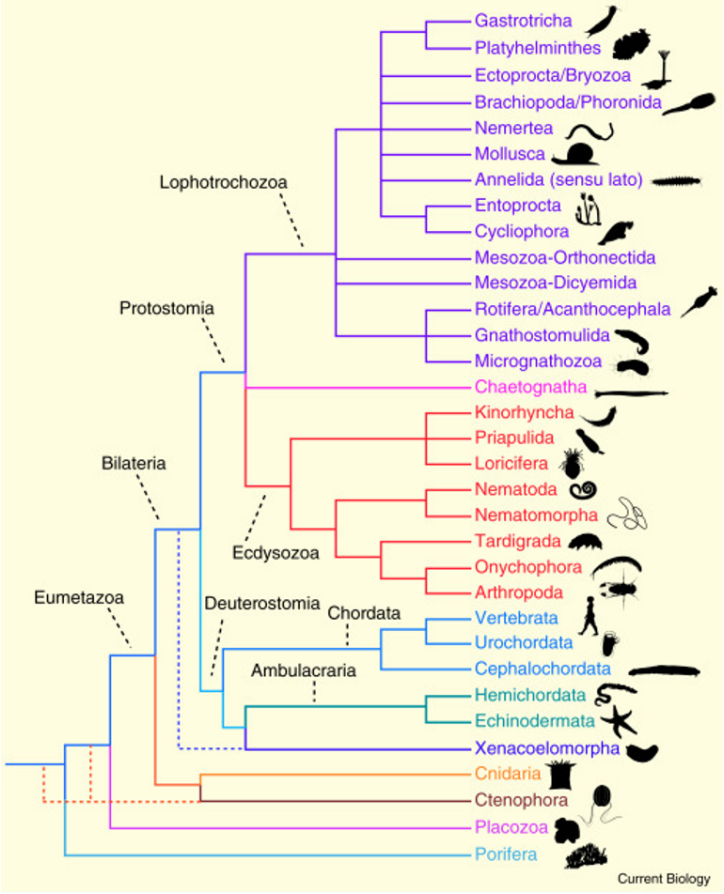


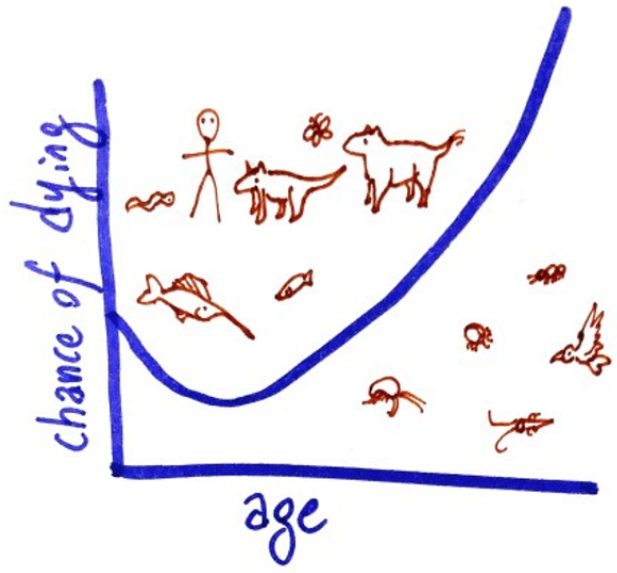
Image from [Telford et al (2015)]

Invertebrates represent most of the strategies that animals have attempted on earth, and certainly most of the animals on earth. *Vertebrates* are the odd ones out.

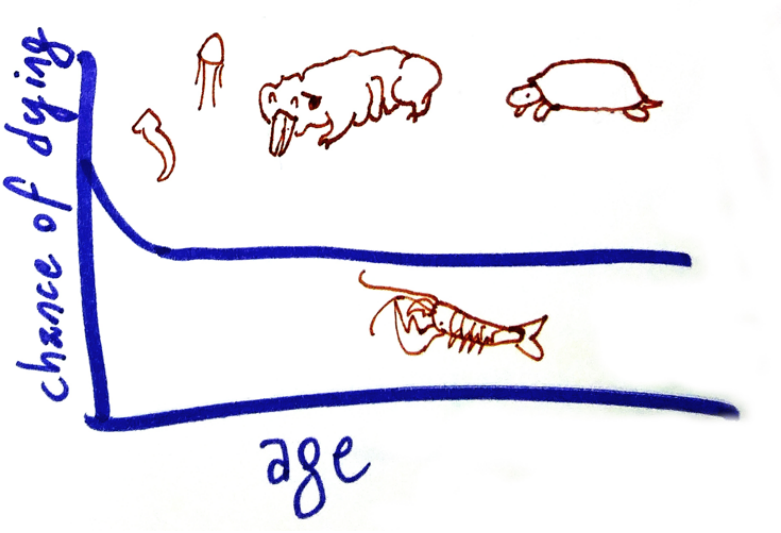
But you know which animals are profoundly weird, no matter which way you look at it? Naked mole rats. Naked mole-rats have like a dozen properties that are not just *unusual*, not just *strange*, but *absolutely batshit*. Let's review.

1. They don't age

What? Well, for most animals, their chance of dying goes up over time. You can look at a population and find something like this:



Mole-rats, they have the same chance of dying at any age. Their graph looks like this:



They’re joined, more or less, by a few species of jellyfish, flatworms, turtles, lobsters, and at least one fish.

They’re hugely long-lived compared to other rodents, seen in zoos at 30+ years old compared to the couple brief years that rats get.

2. They don't get cancer

Cancer generally seems to be the curse of multicellular beings, but naked mole-rats are an exception. A *couple* mole-rats have developed cancer-like growths *in captivity*, but no wild mole-rat has ever been found with cancer.

3. They don’t feel some forms of pain

Mole-rats don’t respond to acid or capsaicin, which is, as far as I know, unique among mammals.

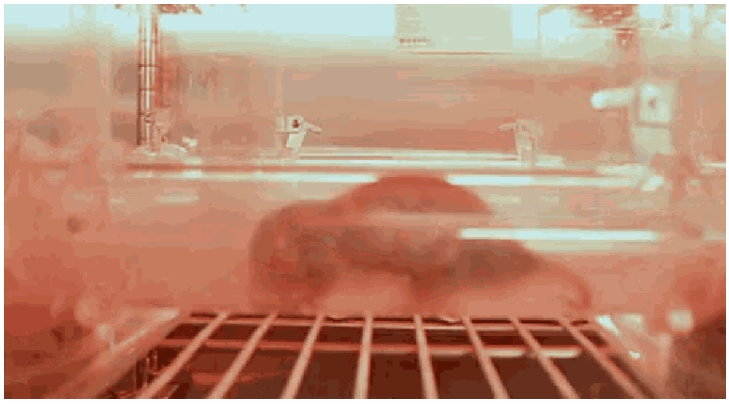
4. They’re eusocial

Definitely unique among mammals. Like bees, ants, and termites, naked mole-rats have a single breeding “queen” in each colony, and other “worker” individuals exist in castes that perform specific tasks. In an evolutionary sense, this means that the “unit of selection” for the species is the queen, not any individual - the queen’s genes are the ones that get passed down.

They’re also a fascinating case study of an animal whose existence was deduced before it was proven. Nobody knew about eusocial mammals for a long time. In 1974, entomologist Richard Alexander, who studied eusocial insects, wrote down a set of environmental characteristics he thought would be required for a eusocial mammal to evolve. Around 1981 and the next decade, naked mole-rats - a perfect match for his predictions - were found to be eusocial.

5. They don’t have fur

Obviously. But aside from genetic flukes or domesticated breeds, that puts them in a small unlikely group with only some marine mammals, rhinoceros, hippos, elephants, one species of boar, and… us.



You and this entity have so much in common.

6. They’re able to survive ridiculously low oxygen levels

It uses very little oxygen during normal metabolism, much less than comparable-sized rodents, and it can survive for hours at 5% oxygen (a quarter of normal levels.)

7. Their front teeth move back and forth like chopsticks

I'm not actually sure how common this is in rodents. But it really weirded me out.

8. They have no regular sleep schedule

This is weird, because *jellyfish* have sleep schedules. But not mole-rats!

9. They're cold-blooded

They have basically no ability to adjust their body temperature internally, perhaps because their caves tend to be rather constant temperatures. If they need to be a different temperature, they can huddle together, or move to a higher or lower level in their burrow.

All of this makes me think that mole-rats must have some underlying unusual properties which lead to all this - a "weirdness generator", if you will.

A lot of these are connected to the fact that mole rats spend almost their entire lives underground. There are lots of burrowing animals, but "almost their entire" is pretty unusual - they don't surface to find food, water, or (usually) mates. (I think they might only surface when digging tunnels and when a colony splits.) So this might explain (8) - no need for a sleep schedule when you can't see the sun. It also seems to explain (5) and (9), because thermoregulation is unnecessary when they're living in an environment that's a pretty constant temperature.

It probably explains (6) because lower burrow levels might have very little oxygen most of the time, although there's some debate about this - their burrows might actually be pretty well ventilated.

And Richard Alexander's 12 postulates that would lead to a eusocial vertebrate - plus some other knowledge of eusociality - suggests that this underground climate, when combined with the available lifestyle and food source of a molerat, should lead to eusociality.

It *might* also be the source of (2) and (3) - people have theorized that higher CO2 or lower oxygen levels in burrows might reduce DNA damage or related to neuron function or something. (This would also explain why only mole-rats in captivity have had tumors, since they're kept at atmospheric oxygen levels.) These still seem to be up in the air, though. Mole-rats clearly have a variety of fascinating biochemical tricks that are still being understood.

So there's at least one "weirdness generator" that leads to all of these strange mole-rat properties. There might be more.

I'm pretty sure it's not the chopstick teeth (7), at least - but as with many predictions one could make about mole rats, I could easily be wrong.

To watch some naked mole-rats going about their lives, check out the Pacific Science Center's mole-rat live camera. It's really fun, if a writhing mass of playful otters that are also uncooked hotdogs sounds fun to you.

Complex Behavior from Simple (Sub)Agents

Epistemic Status: Simultaneously this is work that took me a long time and a lot of thought, and also a playful and highly speculative investigation. Consider taking this seriously but not literally.

Introduction

Take a simple agent (GitHub; Python), with no capacity for learning, that exists on a 2D plane. It shares the plane with other agents and objects, to be described shortly.

The agent intrinsically doesn't want anything. But it can be assigned goal-like objects, which one might view as subagents. Each individual goal-like subagent can possess a simple preference, such as a desire to reach a certain region of space, or a desire to avoid a certain point.

The goal-like subagents can also vary in the degree to which they remain satisfied. Some might be permanently satisfied after achieving their goal once; some might quickly become unsatisfied again after a few timesteps.

Every timestep, the agent considers ten random movements of unit-distance, and executes the movement corresponding to the highest expected valence being reported by its goal-like subagents, in a winner-take-all fashion.

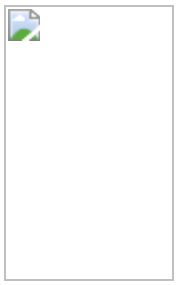
Even with such an intentionally simplistic model, a surprising and illuminating level of behavioral complexity can arise.

Sections 1-8 concern interesting or amusing behaviors exhibited by the model.

Sections 8-12 outline future directions for the model and ruminations on human behavior.

1. Baseline

In this image, the path of the agent is painted with points, the color of the points changing slowly with the passage of time. This agent possesses three subagents with preferences for reaching the three green circles, and a fourth mild preference for avoiding the red circle.



Once it comes within a set distance of one of the green circles, the corresponding subagent is satisfied, and thus the movement with the highest expected valence switches to the next-highest valence goal. The satisfaction gradually wears off, and the agent begins to be drawn to the goal again. Thus, the agent moves inexorably around the triangle of green circles, sometimes in a circuit, sometimes backtracking.

2. "Ugh field"

If the aversion to the red circle is amplified above a certain threshold, this behavior results. The subagent with preferences for reaching the top green circle still exists, but it will never be satisfied, because expected negative valence of passing near the red circle is too high.

But if one is clever, one can find a way around aversions, by inventing intermediary goals or circumventing the aversion with intermediate desirable states.

Sometimes a you want to accomplish something, but a seemingly trivial inconvenience will arise to screen off your motivation. If you can't remove the inconvenience, you can usually find a path around it.

3. Smartphone

What if the agent has, pinned to its position (such that it is constantly somewhat nearby), a low-valence rewarding object, which doesn't provide lasting satisfaction? (In other words - the agent has a goal-like subagent which mildly but relatively persistently wants to approach the pinned object.)

The agent suddenly looks very distracted, doesn't it? It doesn't make the same regular productive circuits of its goals. It seems to frequently get stuck, sphexishly returning to a goal that it just accomplished, and to take odd pointless energy-wasting zigzags in its path.

Maybe it's bad to constantly carry around attention-grabbing objects that provide us with miniscule, unsatisfying hits of positive valence.

Considered together, Parts 2 and 3 speak to the dangers of convenience and the power of trivial inconvenience. The agents (and humans) are extraordinarily sensitive not only to the absolute valence of an expectation, but to the proximity of that state. Even objectively weak subagents can motivate behavior if they are unceasingly present.

4. Agitation and/or Energy

The model does not actually have any concept of energy, but it is straightforward to encode a preference for moving around a lot. When the agent is so inclined, its behavior becomes chaotic.

Even a relatively moderate preference for increased movement will lead to some erratic swerves in behavior.

If one wished, one could map this type of behavior onto agitation, or ADHD, or anxiety, or being overly caffeinated. On the other hand, you could view some degree of "restlessness" as a drive toward exploration, without which one might never discover new goals.

One path of investigation that occurred to me but which I did not explore was to give the agent a level of movement-preference that waxed and waned cyclically over time. Sometimes you subjectively have a lot of willpower, sometimes you subjectively *can't* focus on anything. But, on the whole, we all manage to get stuff done.

5. Look-Ahead

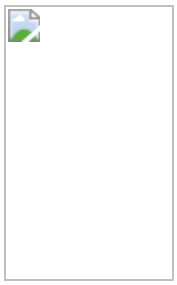
I attempted to implement an ability for the agent to scan ahead more than one step into the future and take the movement corresponding the highest expected valence in *two* timesteps, rather than just the next timestep. This didn't really show anything interesting, and remains in the category of things that I will continue to look into. (The Red agent is thinking two moves ahead, the Blue agent only one move ahead. Is there a difference? Is the clustering of the Red agent's pathing slightly tighter? Difficult to say.)

I don't personally think humans explicitly look ahead very often. We give ourselves credit as the "thinking, planning animal", but we generally just make whichever choice corresponds to the highest expected valence in the current moment. Looking ahead is also very computationally expensive - both for people, and for these agents - because it inevitably requires something like a model-based tree search. What I think we *actually* do is better addressed in Section 10 regarding Goal Hierarchies.

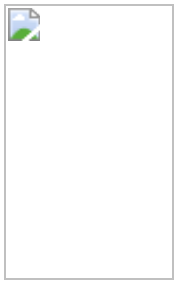
6. Sociability

Of course, we can give the agents preferences for being near other agents, obeying the same rules as the preferences for any other position in space.

With hyper-dominant, non-extinguishing preferences for being around other agents, we get this piece of computer generated art that I call "Lovers".



With more modest preference for the company of other agents, and with partially-overlapping goals (Blue agent wants to spend time around the top and rightmost target, Red agent wants to spend time around the top and leftmost target) you get this *other* piece of art that I call "Healthy Friendship". It looks like they're having fun, doesn't it?



7. New Goals Are Disruptive

Brief reflection should confirm that introducing a new goal into your life can be very disruptive to your existing goals. You could say that permitting a new goal-like subagent to take root in your mind is akin to introducing a competitor who will now be bidding against all your existing goals for the scarce resource of your time and attention.

Compare this image with the Baseline at the top of this article. The new, powerful top-right goal has siphoned away all the attention from the formerly stable, well-tended trio of goals.

I think one of the main reasons we fall down on our goals is simply that we spontaneously generate new goals, and these new goals disrupt our existing motivational patterns.

8. Winner-Take-All?

You may have more questions about the winner-take-all assumption that I mentioned above. In this simple model, the goal-like subagents do not "team up". If two subagents would prefer that the agent move to the left, this does not mean that their associated valence will sum up and make that choice more globally appealing. The reason is simple: if you straightforwardly sum up over all valences instead of picking a winner, this is what happens:

The agent simply seeks out a local minimum and stays there.

I am currently somewhat agnostic as to what the human or animal brain is actually doing. We do appear to get stuck in local minima sometimes. But you can get sphexish behavior that looks like a local minimum out of a particular arrangement of winner-take-all subagents. For example, if an agent is hemmed in by aversive stimuli with no sufficiently positive goal-states nearby, that might look like a local minimum, though it is still reacting to each aversive stimulus in a winner-take-all fashion.

Subjectively, though, it feels like if you have two good reasons supporting an action, that makes the action feel a bit easier to do, a bit more motivating, than if you just had one good reason. This hints that maybe goal-like subagents can gang up together. But I also doubt that this is anything like strictly additive. Thinking of 2,000 reasons why I should go to the gym isn't 2,000 times more compelling than thinking of one reason.

9. Belief, Bias, and Learning

The main area of the model that I would like to improve, but which would amplify the complexity of the code tremendously, would be in introducing the concept of bias and/or belief. The agent should be able to be *wrong* about its expected valence. I think this is hugely important, actually, and explains a lot about human behavior.

Pathologies arise when we are systematically wrong about how good, or how bad, some future state will be. But we can overcome pathologies by exposing ourselves to those states, and becoming deeply calibrated regarding their reality. On the aversion side this applies to everything from the treatment of phobias and PTSD, to the proper response to a reasonable-seeming anxiety. On the positive-valence side, we may imagine that it would be incredibly cool and awesome to do or to be some particular thing, and only experience can show us that accomplishing such things yields only a shadow of what we expected. Then your brain updates on that, and you cease to feel

motivated to do that thing anymore. You can no longer sustain the delusion that it was going to be awesome.

10. Goal Hierarchies

It seems clear that, in humans, goals are arranged in something like trees: I finish this current push-up because I want to finish my workout. I want to finish my workout because I want to stay on my workout program. I want to stay on my workout program because I want to be strong and healthy.

But it's almost certainly more complex than this, and I don't know how the brain manages its "expected valence" calculations across levels of the tree.

I hypothesize that it goes something like this. Goal-like subagents concerned with far-future outcomes, like "being strong and healthy", generate (or perhaps *manifest as*) more specific near-term goal-like targets, with accompanying concrete sensory-expectation targets, like "working out today". This seems like one of those mostly automatic things that happens whether or not we engineer it. The automaticity of it seems to rely on our maps/models/beliefs about how the world works. Even much simpler animals can chain together and break down goals, in the course of moving across terrain toward prey, for example.

The model described above doesn't really have a world model and can't learn. I could artificially designate some goals as being sub-goals of other goals, but I don't think this is how it actually works, and I don't actually think it would yield any more interesting behavior. But it might be worth looking into. Perhaps the most compelling aspect of this area is that what would be needed would not be to amplify the cleverness of the agent; it would be to amplify the cleverness of the *subagent* in manipulating and making its preferences clearer to the agent. For example: give subagents the power to generate new goal-objects, and lend part of their own valence to those subagents.

11. Suffering

I toyed with the idea of summing up all the valences of the goal-objects that were being ignored at any given moment, and calling that "suffering". This sure is what suffering feels like, and it's akin to what those of a spiritual bent would call suffering. Basically, suffering is wanting contradictory, mutually exclusive things, or, being aware of wanting things to be a certain way while simultaneously being aware of your inability to work toward making it that way. One subagent wants to move left, one subagent wants to move right, but the agent has to pick one. Suffering is something like the expected valence of the subagent that is left frustrated.

I had a notion here that I could stochastically introduce a new goal that would minimize total suffering over an agent's life-history. I tried this, and the most stable solution turned out to be thus: introduce an overwhelmingly aversive goal that causes the agent to run far away from all of its other goals screaming. Fleeing in perpetual terror, it will be too far away from its attractor-goals to feel much expected valence towards them, and thus won't feel too much regret about running away from them. And it is in a sense satisfied that it is always getting further and further away from the object of its dread.

File this under "degenerate solutions that an unfriendly AI would probably come up with to improve your life."

I think a more well-thought-out definition of suffering might yield much more interesting solutions to the suffering-minimization problem. This is another part of the model I would like to improve.

12. Happiness and Utility

Consider our simple agents. What makes them happy?

You could say that something like satisfaction arises the moment they trigger a goal-state. But that goal object immediately begins recharging, becoming "dissatisfied" again. The agent is never actually content, unless you set up the inputs such that the goal valences don't regenerate - or if you don't give it goals in the first place. But if you did that, the agent would just wander around randomly after accomplishing its goals. That doesn't seem like happiness.

Obviously this code doesn't experience happiness, but when I look at the behavior of the agents under different assumptions, the agents *seem* happy when they are engaged in accomplishing their various goals. They seem *unhappy* when I create situations that impede the efficiency of their work. This is obviously pure projection, and says more about me, the human, than it says about the agent.

So maybe a more interesting question: What are the high-utility states for the agent? At any given moment of time the agents certainly have preference orderings, but those preference orderings shift quite dramatically based on its location and the exact states of each of its subagents, specifically their current level of satisfaction. In other words, in order to mathematically model the preference ordering of the agent across all times, you must model the individual subagents.

If humans "actually" "have" "subagents" - whatever those words actually end up meaning - then the "human utility function" will need to encompass each and every subagent. Even, I think, the very stupid ones that you don't reflectively endorse.

Conclusion

I set out on this little project because I wanted to prove some assumptions about the "subagent" model of human consciousness. I don't think I can ultimately say that I "proved" anything, and I'm not sure that one could ever "prove" anything about human psychology using this particular methodology.

The line of thinking that prompted this exploration owes a lot to Kaj_Sotala, for his ongoing Sequence, Scott Alexander's reflections on motivation, and Mark Lippman's Folding material. It's also their fault I used the unwieldy language "goal-like subagent" instead of just saying "the agent has several goals". I think it's much more accurate, and useful, to think of the mind as being composed of subagents, than to say it "has goals". Do you "have" goals if the goals control you?

This exercise has changed my inner model of my own motivational system. If you think long enough in terms of subagents, something eventually clicks. Your inner life, and your behaviors, seems to make a lot more sense. Sometimes you can even leverage this perspective to construct better goals, or to understand where some goals are actually coming from.

The code linked at the top of this page will generate all of the figures in this article. It is not especially well documented, and bears the marks of having been programmed by a feral programmer raised in the wilds of various academic and industrial

institutions. Be the at as it may, the interface is not overly complex. Please let me know if anyone ends up playing with the code and getting anything interesting out of it.

Say Wrong Things

There are many ways you might approach being less wrong.

A popular one is to make fewer wrong statements; to say fewer wrong things.

Naively it would seem this is a recipe for success, since you just say more things that are true and right and fewer things that are false and wrong. But if Goodhart has anything to say about it, and he does, you'll find ways to maximize the measure at the expense of the original objective.

Assuming the real objective is something like "have a more complete, precise, and accurate model of the world that better predicts the outcome of subjectively unknown events", then we can quickly see the many ways Goodharting can lead us astray if we focus too much on appearing less wrong. We might:

- make fewer claims than we could, pulling us away from completeness even as we appear less wrong;
- make weaker claims than we could, pulling us away from precision;
- and, a perennial favorite, filter the claims we publicly make so we appear less wrong than we really are by hiding our least confident claims.

The first two can be corrected with better calibration, that is by making statements with confidence intervals or likelihoods that proportionally match the observed frequency of correctness of similarly confident claims. But simply suggesting someone "be better calibrated" is not a motion they can make; it's an outcome of taking actions towards increasing calibration. As good a place to start as any for improving calibration is the forecasting literature, if that's what you'd like to do.

The third is more tricky, though, because it's less directly about claims being made and their probability of correctness and more about social dynamics and how you appear to other people. And for that reason it's what I want to focus on here.

Appearing Less Wrong

I've met a lot of people in my life who are experts at not looking as stupid as they are.

That's kind of harsh. Maybe a nicer way to say it is that they are experts at appearing to be better at making correct predictions about the world than they actually are.

Some of their techniques are just normal social tricks: projecting confidence, using social status, the ever-abused term "gaslighting", and other methods of getting people to believe they are right even when a more careful examination would reveal them to be mistaken. These are people we all love to hate and love when we can call them on their bullshit: overconfident academics, inflated politicians, self-important internet intellectuals, and those people whose idea of social interaction is to say "well, actually...".

But there's a way to avoid looking stupid that is more pernicious, less amenable to calling out, and that subtly drags you towards local maxima that trap you mountains

and valleys away from more complete understanding. And it's to **shut up and not tell anyone about your low confidence beliefs**.

It is extremely tempting to do this. Among the many benefits of keeping low probability claims to yourself:

- you have a high accuracy ratio of publicly made claims, making you look right more often when observed;
- you say only things that, even when wrong, turn out to be wrong in conservative ways that still make you look smart;
- and you accrue a reputation of being right, usually conferring social status, which can feel really good.

The only trouble is that this approach is too conservative, too modest. It's easy to justify this kind of outward modesty as keeping up appearances in a way that is instrumental to some bigger goal, and you say to yourself "I'll still make low probability claims; I'll just keep them to myself", but down that path lies shadow rationality via compartmentalization. You can try it, but good luck, because it's a dark art that hopes to do what human brains cannot, or at least cannot without some sufficiently powerful magic, and that magic traditionally comes with vows not to do it.

Meanwhile, out in the light, finding models that are better predictive of reality sometimes requires holding beliefs that appear unlikely to be true but then turn out to be right, sometimes spectacularly so, although *semper caveat*, all models are wrong, but some are useful. And then you have to go all in sometimes, exploring the possibility that your 10% guess turns out to be 100% correct, minus epsilon, because if you don't do this you'll do no better than the medieval Scholastic holding to Aristotelian physics or the early 20th century geologist ignoring the evidence for continental drift, forever locked away from taking the big risks necessary to find better, more accurate, precise, and complete models.

Okay, so let's say you are convinced not to try so hard to appear more right than you are. How do you do that?

Say It Wrong

So I suppose it's nothing so much harder than just telling people your claims, even if you have low confidence in them, and seeing how they react, although depending on the circumstances you'll probably want to adequately explain your confidence level so they can update on it appropriately. The trouble is getting yourself to do that.

I can't change your mind for you, although thankfully some folks have developed some techniques that might help if you're not interested in over-solving that problem. What I can do is point out a few things that might help you see where you are being too modest, nudge you towards less modesty, and create an environment where it's safe to be less modest.

- Look for the feeling of "pulling your punches" when you are telling people your ideas.
- Ramble more and filter less.
- Alternatively, more babble less prune.
- Worry less about how you look to others.

- Relatedly, increase your ability to generate your own esteem so you less need to receive it from others, giving you more freedom to make mistakes.
- Encourage others to tell you their half-baked ideas.
- When they do, be supportive.
- Take a collaborative, nurturing approach to truth seeking.
- Play party games like "what do you think is true that you think no one else here also thinks is true?" and "what's your least popular opinion?".
- And my personal favorite, write up and share your immodest, lower-confidence ideas that you think deserve exploration because they have high expected value if they turn out to be right.

I suspect that the real reason most people try too hard to appear more right than they are is fear—fear of being wrong, fear of looking stupid, fear of losing status, fear of losing prestige, fear of losing face, fear of being ostracized, fear of being ignored, fear of feeling bad, fear of feeling lesser. I see this fear, and I honor it, but it must be overcome if one wishes to become stronger. And when you fear being more wrong, you will be too careful to ever become as less wrong as you could.

To sum it all up pithily:

To become less wrong, you must give up being most right.

Quotes from Moral Mazes

Reading and actually paying attention to Moral Mazes is hard. Writing carefully about it is even harder. I effectively spent several months forcing my way through the book, because it seemed important to do that. I then spent a month trying to write about the book, but that's going super slow as well. The repetition, the saying the same thing from multiple angles, the detailed examples, seem necessary to get the points across, because one has the very strong instinct to avoid understanding it, to read without seeing, to hear without listening. At least, I know I did, despite these things also not only not feeling new, but resonating with my direct experiences.

So in the interest of getting something out there, and hoping that I'll be able to address things in more detail later, here are all the 168 (!) quotes I highlighted from the book, roughly organized into categories. Locations listed are how to find the quote in the Kindle edition, and the quotes are numbered for ease of search and reference.

A. Hierarchy and Credit

1. As a former vice-president of a large firm says: "What is right in the corporation is not what is right in a man's home or in his church. What is right in the corporation is what the guy above you wants from you. That's what morality is in the corporation." (Location 148)

2. When managers describe their work to an outsider, they almost always first say: "I work for [Bill James]" or "I report to [Harry Mills]" or "I'm in [Joe Bell's] group,"* and only then proceed to describe their actual work functions. (Location 387)

3. The key interlocking mechanism of this structure is its reporting system. Each manager gathers up the profit targets or other objectives of his or her subordinates and, with these, formulates his commitments to his boss; this boss takes these commitments and those of his other subordinates, and in turn makes a commitment to his boss.* (Location 441)

4. It is characteristic of this authority system that details are pushed down and credit is pulled up. (Location 446)

5. One gives credit, therefore, not necessarily where it is due, although one always invokes this old saw, but where prudence dictates. Customarily, people who had nothing to do with the success of a project can be allocated credit for their exemplary efforts. At the middle levels, therefore, credit for a particular idea or success is always a type of refracted social honor; one cannot claim credit even if it is earned. Credit has to be given, and acceptance of the gift implicitly involves a reaffirmation and strengthening of fealty. A superior may share some credit with subordinates in order to deepen fealty relationships and induce greater efforts on his behalf. Of course, a different system obtains in the allocation of blame. (Location 482)

6. If one has a mistake-prone boss, there is, of course, always the temptation to let him make a fool of himself, but the wise subordinate knows that this carries two dangers—he himself may get done in by his boss's errors, and, perhaps more important, other managers will view with the gravest suspicion a subordinate who withholds crucial information from his boss even if they think the boss is a nincompoop. A subordinate must also not circumvent his boss nor ever give the

appearance of doing so. He must never contradict his boss's judgment in public. To violate the last admonition is thought to constitute a kind of death wish in business, and one who does so should practice what one executive calls "flexibility drills," an exercise "where you put your head between your legs and kiss your ass good-bye." (Location 424)

7. The general rule is that bosses are expected to protect those in their bailiwicks. Not to do so, or to be unable to do so, is taken as a sign of untrustworthiness or weakness. If, however, subordinates make mistakes that are thought to be dumb, or especially if they violate fealty obligations—for example, going around their boss—then abandonment of them to the vagaries of organizational forces is quite acceptable. (Location 438)

8. Managers often note that one must stay at least three drinks behind one's boss at social functions; this meant that Brown's subordinates might never drink at all on such occasions. (Location 630)

9. However, the belief of insiders in abstract goals is not a prerequisite for personal success; belief in and subordination to individuals who articulate organizational goals is. One must, however, to be successful in a bureaucratic work situation, be able to act, at a moment's notice, as if official reality is the only reality. (Location 1194)

10. You can put the damper on anyone who works for you very easily and that's why there's too much chemistry in the corporation. There's not enough objective information about people. When you really want to do somebody in, you just say, well, he can't get along with people. That's a big one. And we do that constantly. What that means, by the way, is that he pissed me off; he gave evidence of his frustration with some situation. Another big one is that he can't manage—he doesn't delegate or he doesn't make his subordinates keep his commitments. So in this sort of way, a consensus does build up about a person and a guy can be dead and not even know it. (Location 1475)

11. Only at this point did Brady realize that it was the CEO himself who was fiddling with the numbers. The entire dual reporting system that the CEO had personally initiated was in part an elaborate spy network to guard against discovery of the slush fund manipulation, and perhaps other finagling, rather than a system to ensure financial honesty. (Location 2404) [Leads into next quote]

12. [Note: Brady is an accountant.] The corporate managers to whom I presented this case see Brady's dilemma as devoid of moral or ethical content. In their view, the issues that Brady raises are, first of all, simply practical matters. His basic failing was, first, that he violated the fundamental rules of bureaucratic life. These are usually stated briefly as a series of admonitions. (1) You never go around your boss. (2) You tell your boss what he wants to hear, even when your boss claims that he wants dissenting views. (3) If your boss wants something dropped, you drop it. (4) You are sensitive to your boss's wishes so that you anticipate what he wants; you don't force him, in other words, to act as boss. (5) Your job is not to report something that your boss does not want reported, but rather to cover it up. You do what your job requires, and you keep your mouth shut. Second, the managers that I interviewed feel that Brady had plenty of available legitimations to excuse or justify his not acting. Clearly, they feel, a great many other executives knew about the pension fund scam and did nothing; everybody, especially the top bosses, was playing the game. The problem fell into other people's areas, was their responsibility, and therefore their problem. Why, then, worry about it? Besides, Brady had a number of ways out of the situation if he

found it intolerable, including resigning. Moreover, whatever action he took would be insignificant anyway so why bother to act at all and jeopardize himself? Even a fool should have known that the CEO was not likely to take whatever blame resulted from the whole affair. (Location 2429)

13. One's best efforts at being fair, equitable, and generous with subordinates clash both with a logic that demands choices between people, inevitably producing hatred, envy and animosity, and with the plain fact that, despite protestations to the contrary, many people do not want to be treated fairly. (Location 4552)

14. Mastering the subtle but necessary arts of deference without seeming to be deferential, of "brown nosing" without fawning, of simultaneous self-promotion and self-effacement, and occasionally of the outright self-abasement that such relationships require is a taxing endeavor that demands continual compromises with conventional and popular notions of integrity. Only those with an inexhaustible capacity for self-rationalization, fueled by boundless ambition, can escape the discomfort such compromises produce. (Location 4572)

B. Feeling Comfortable

15. Essentially, managers try to gauge whether they feel "comfortable" with proposed resolutions to specific problems, a task that always involves an assessment of others' organizational morality and a reckoning of the practical organizational and market exigencies at hand. The notion of comfort has many meanings. When applied to other persons, the idea of comfort is an intuitive measure of trustworthiness, reliability, and predictability in a polycentric world that managers often find troubling, ambiguous, and anxiety-laden. Such assessment of others' organizational morality is a crucial aspect of a more general set of probations that are intrinsic to managerial work. (Location 302)

16. They objected in particular to those aspects of my brief written proposal that discussed the ethical dilemmas of managerial work. They urged me to avoid any mention of ethics or values altogether and concentrate instead on the "decision-making process" where I could talk about "trade-offs" and focus on the "hard decisions between competing interests" that mark managerial work. Taking these cues, I rewrote and rewrote the proposal couching my problem in the bland, euphemistic language that I was rapidly learning is the lingua franca of the corporate world. But such recasting eroded whatever was distinctive about the project and some managers dismissed the study as a reinvention of the wheel. (Location 324)

17. In effect, I could not get access to study managers' moral rules-in-use because I seemed unable to articulate the appropriate stance that would convince key managers that I already understood those rules and was thus a person with whom they could "feel comfortable" enough to trust. (Location 334)

18. The process centered on the written proposal that I had been circulating and consisted essentially of a furthering of my linguistic education in the art of indirect rather than pointed statement and, more particularly, a reformulation of my inquiry that recast the moral issues of managerial work as issues of public relations. When, after several rewritings, the proposal satisfied him, he approached a well-placed executive in a large textile firm that I have given the pseudonym of Weft Corporation and vouched for me. At that point, the proposal itself became meaningless since, to my knowledge, no one except the two executives who arranged access ever saw it. The personal vouching, however, was crucial. This was based on what both men took

to be a demonstrated willingness and ability to be “flexible” and especially on their perception that I already grasped the most salient aspect of managerial morality as managers themselves see it—that is, how their values and ethics appear in the public eye. (Location 347)

19. At bottom, all of the social contexts of the managerial world seek to discover if one “can feel comfortable” with another manager, if he is someone who “can be trusted,” if he is “our kind of guy,” or, in short, if he is “one of the gang.” (Location 905)

20. My search for access involved me in some of the crucial bureaucratic intricacies that shape managers’ experiences. These include organizational upheavals, political rivalries, linguistic ambiguity, the supremacy of chance and tangled personal connections over any notion of intrinsic merit, the central significance of public relations, and, perhaps especially, the ceaseless moral probations for inclusion in a managerial circle. Managers keep their eyes on the organizational premiums that shape behavior, values, ethics, and worldviews in corporate bureaucracies. I focus on those premiums… (Location 387)

21. One becomes known, for instance, as a trusted friend of a friend; thought of as a person to whom one can safely refer a thorny problem; considered a “sensible” or “reasonable” or, especially, a “flexible” person, not a “renegade” or a “loose cannon rolling around the lawn”; known to be a discreet person attuned to the nuances of corporate etiquette, one who can keep one’s mouth shut or who can look away and pretend to notice nothing; or considered a person with sharp ideas that break deadlocks but who does not object to the ideas being appropriated by superiors. (Location 870)

22. Similarly, Covenant’s CEO sold large tracts of land with valuable minerals at dumbfoundingly low prices. The CEO and his aides said that Covenant simply did not have the experience to mine these minerals efficiently, a self-evident fact from the low profit rate of the business. In all likelihood, according to a manager close to the situation, the CEO, a man with a financial bent and a ready eye for the quick paper deal, felt so uncomfortable with the exigencies of mining these minerals that he ignored the fact that the prices the corporation was getting for the minerals had been negotiated forty years earlier. Such impulsiveness and indeed, one might say from a certain perspective, irrationality, is, of course, always justified in rational and reasonable terms. It is so commonplace in the corporate world that many managers expect whatever ordered processes they do erect to be subverted or overturned by executive fiat, masquerading as an established bureaucratic procedure or considered judgment. (Location 1700)

23. think that I’ve got to where I am today because of this. [His boss’s boss] knows that I saved the company a lot of money and a lot of asses to boot. And he and others know that I am someone who can be trusted. I can keep my mouth shut…. And that’s the biggest thing that I have going for me—that people feel that I can be trusted. I can’t overemphasize that enough. (Location 2917)

24. Only those men and women who allow peers and superiors to feel morally comfortable in the ambiguous muddles of the world of affairs have a chance to survive and flourish in big organizations when power and authority shift due to changes in markets, internal power struggles, or the need to respond to external exigencies. (Location 4943)

C. Struggle for Success

25. The logical result of alertness to expediency is the elimination of any ethical lines at all. (Location 2985)

26. The two areas are, of course, related since one's chances in an organization depend largely on one's "credibility," that is, on the widespread belief that one can act effectively. One must therefore prevail regularly, though not always, in small things to have any hope of positioning oneself for big issues. The hidden agenda of seemingly petty disputes may be a struggle over long-term organizational fates. (Location 812)

27. A fundamental rule of corporate politics is that one never cedes control over assets, even if the assets are administrative headaches. (Location 643)

28. Bureaucratic hierarchies, simply by offering ascertainable rewards for certain behavior, fuel the ambition of those men and women ready to subject themselves to the discipline of external exigencies and of their organization's institutional logic, the socially constructed, shared understanding of how their world works. However, since rewards are always scarce, bureaucracies necessarily pit people against each other and inevitably thwart the ambitions of some. (Location 806)

29. When asked who gets ahead, an executive vice-president at Weft Corporation says: The guys who want it [get ahead]. The guys who work. You can spot it in the first six months. They work hard, they come to work earlier, they leave later. They have suggestions at meetings. They come into a business and the business picks right up. They don't go on coffee breaks down here [in the basement]. You see the parade of people going back and forth down here? There's no reason for that. I never did that. If you need coffee, you can have it at your desk. Some people put in time and some people work. (Location 992)

30. Proper management of one's external appearances simply signals to one's peers and to one's superiors that one is prepared to undertake other kinds of self-adaptation. Managers also stress the need to exercise iron self-control and to have the ability to mask all emotion and intention behind bland, smiling, and agreeable public faces. (Location 1059)

31. The price of bureaucratic power is a relentlessly methodical subjection of one's impulses, at least in public. (Location 1100)

32. [Style] is being able to talk easily and make presentations. To become credible easily and quickly. You can advance quickly even without technical experience if you have style. You get a lot of points for style. You've got to be able to articulate problems, plans, and strategies without seeming to have to refer to all sorts of memos and so on. The key in public performances and presentations is in knowing how to talk forcefully without referring to notes and memoranda. To be able to map out plans quickly and surely. (Location 1298)

33. As one manager says: "Personality spells success or failure, not what you do on the field." (Location 1383)

34. More generally, there are several rules that apply here. First, no one in a line position—that is, with responsibility for profit and loss—who regularly "misses his numbers" will survive, let alone rise. Second, a person who always hits his numbers

but who lacks some or all of the required social skills will not rise. Third, a person who sometimes misses his numbers but who has all the desirable social traits will rise. (Location 1412)

35. A managerial commonplace says: “In the corporate world, 1,000 ‘Attaboys’ are wiped away with one ‘Oh, shit!’” (Location 1619)

36. There’s a lot of it [fear and anxiety]. To a large degree it’s because people are more honest with themselves than you might believe. People know their own shortcomings. They know when they’re over their heads. A lot of people are sitting in jobs that they know are bigger than they should be in. But they can’t admit that in public and, at still another level, to themselves. The organizational push for advancement produces many people who get in over their heads and don’t know what they are doing. And they are very fearful of making a mistake and this leads to all sorts of personal disloyalty. But people know their capabilities and know that they are on thin ice. And they know that if they make mistakes, it will cost them dearly. So there’s no honesty in our daily interaction and there’s doubt about our abilities. The two go together. (Location 1767)

37. One comes to gauge that hard-won access to managerial circles takes precedence over fussing with abstract principles. (Location 2923)

38. Perceptions of pervasive mediocrity breed an endless quest for social distinctions even of a minor sort that might give one an “edge,” enable one to “step out of the crowd,” or at least serve as a basis for individual claims to privilege. More specifically, an atmosphere of mediocrity erodes the hope of meaningful collective achievement and encourages, at least among more aggressive managers, a predatory stance toward their organizations, that is, a search for private deals, a working of the system for one’s own personal advantage. (Location 4436)

39. One leaves behind as well the technical knowledge or scientific expertise of one’s younger years, lore now more suited for the narrower roles of technicians or junior managers. One must, in fact, put distance between oneself and technical details of every sort or risk the inevitable entrapment of the particular. Salesmen, too, must leave their bags and regular customers and long boisterous evenings that seal measurable deals behind them and turn to marketing strategies. Work becomes more ambiguous, directed as it is toward maneuvering money, symbols, organizational structures, and especially people. The CEO at Weft Corporation, it is said, “doesn’t know a loom from a car.” And the higher one goes, the more managers find that “the essence of managerial work is cronyism, covering your ass, [and] pyramiding to protect your buddies.” (Location 4539)

40. the rewards of corporate success can be very great. And those who do succeed, those who find their way out of the crowded, twisting corridors and into the back rooms where the real action is, where the big games take place, and where everyone present is a player, shape, in a decisive way, the moral rules-in-use that filter down through their organizations. The ethos that they fashion turns principles into guidelines, ethics into etiquette, values into tastes, personal responsibility into an adroitness at public relations, and notions of truth into credibility. (Location 4603)

41. Instead, success becomes contingent on others’ interpretations of one’s performance, leading to a break in the accepted moral economy between talent combined with effort and reward. This makes compulsive sociability an occupational

virtue, as one attempts to discern and shape peers' and superiors' interpretations. (Location 4950)

D. Nothing Matters, But Hit Your Numbers

42. Managers rarely speak of objective criteria for achieving success because once certain crucial points in one's career are passed, success and failure seem to have little to do with one's accomplishments. (Location 917)

43. Corporations rely on other institutions—principally the schools—to establish what might be called competence hurdles. The demonstrated ability of a student to leap over successively higher hurdles in school is taken as evidence of the ability to weather well the probationary trials of corporate life. (Location 920)

44. In Alchemy Inc., whether in sales, marketing, manufacturing, or finance, the "breaking point" in the hierarchy is generally thought to be grade 13 out of 25 or the top 8.5 percent of management. By the time managers reach such a numbered grade in an ordered hierarchy—and the grade is socially defined and varies from company to company—managerial competence as such is taken for granted and assumed not to differ greatly from one manager to the next. (Location 943)

45. A product manager in the chemical company talks about the lack of connection between work and results: I guess the most anxiety provoking thing about working in business is that you are judged on results whether those results are your fault or not. So you can get a guy who has tried really hard but disaster strikes; and you can get a guy who does nothing and his business makes a big success. And so you just never know which way things are going to go and you're never sure about the relationship of your work to the outcome. One of the top executives in Weft Corporation echoes this sentiment: I always say that there is no such thing as a marketing genius; there are only great markets. (Location 1585)

46. Assuming a basic level of corporate resources and managerial know-how, real economic outcome is seen to depend on factors largely beyond organizational or personal control. (Location 1592)

47. upper-middle level manager says: If I were just out of school and somebody told me that it doesn't matter what you do and how well you do it but that what matters is being in the right place at the right time, I'd have said that hard work is still the key. You know, the old virtues. But now as I have gotten older, I think it's pure happenstance—luck. Things happen to people and being in the right time and place and knowing the right people is the key. (Location 1621)

48. It is interesting to note in this context that a line manager's long-run credibility suffers just as much from missing his numbers on the up side (that is, achieving profits higher than predicted) as from missing them on the down side although, as one might expect, the immediate consequences of such different miscalculations vary. Both outcomes, however, undercut the ideology of managerial planning and control. (Location 1595)

49. A top staff official at Covenant Corporation explains: By putting the money in this business, you're taking the money away from others. In human terms, that's what you're doing. It's money that you could provide jobs with to others. So when you get a guy in the business who comes in under or over the plan, well, both are equally suspect. Because you're making major decisions based on your plan…. Like when we

shut down [business A] and put the money into [business B], the whole legitimacy of the operation depends on the [business A] guys accepting the rationale that more money can be made in another operation. (Location 1599)

50. of a plant manager who, when his machinery had ground to a halt and his technicians were baffled and everyone turned to him to make a decision, told his crew, without the faintest idea of the right thing to do and with the great fear that all he had worked for was about to crumble before him, to dump ten pounds of phosphate into the machine. The machine sprang to life and he became a hero. (Location 1656)

E. Implicitness

51. If I tell someone what to do—like do A, B, or C—the inference and implication is that he will succeed in accomplishing the objective. Now, if he doesn't succeed, that means that I have invested part of myself in his work and I lose any right I have to chew his ass out if he doesn't succeed. If I tell you what to do, I can't bawl you out if things don't work. And this is why a lot of bosses don't give explicit directions. They just give a statement of objectives, and then they can criticize subordinates who fail to make their goals. (Location 454)

52. A typical example occurred in Weft Corporation a few years ago when the CEO, new at the time, expressed mild concern about the rising operating costs of the company's fleet of rented cars. The following day, a stringent system for monitoring mileage replaced the previous casual practice. (Location 490)

53. One must remember, for instance, that in our litigious age the best rule in dealing with angry subordinates is to say nothing or as little as possible since whatever one says may be used against oneself and one's organization. (Location 1066)

54. Well, usually you don't tell people the truth. I once knew a guy whom I knew was about to be fired and I asked if he had been told and he had never been told. I think you should tell people explicitly. Things like that shouldn't have to be decoded. But you can understand how it happens. Suppose you have a guy and the consensus is that he isn't promotable. You wouldn't ever—or very seldom—tell him. He goes on to justify his silence: There are people who go through life thinking they can do a lot more than they really can do. And the reason is that losing or changing jobs is a very high stress situation and most people prefer to hang on to what they've got—to their routine. They're not happy but they go through life like prisoners of war not recognizing their true situation. (Location 1494)

55. You get the situation where a lot of people don't really want to know…. Like one guy we have, he will retire on his job. He's in my division. He knows it. I know it. And he doesn't want me to tell him about that. Now don't ask me how I know that but, believe me, I do. (Location 1502)

56. Why does it happen? Because people are afraid of confrontations. People want to be thought of as kind, sensitive, and compassionate. Being compassionate has a good significance in our society. The easy way out is not to do anything, don't tell the guy. That happens a lot. (Location 1510)

57. As a matter of fact, he wouldn't even have to say "cut capital." He would just put pressure on him by saying: "Well, sales are down 50 percent; why aren't your

expenses down 50 percent?” My boss will come to me, by the time it reaches him, and say: “Cut costs.” It’s as simple as that. (Location 2078)

58. He put things into writing in a world that, apart from ritual nods to the importance of documentation, actually fosters ambiguity by its reliance on talk as the basic mode of negotiation and command. Talk, of course, lends itself more readily than documents to backtracking, filling in, evasion, subterfuge, and secrecy, all important virtues if one is to do what has to be done while establishing and maintaining the kinds of relationships that alone can protect oneself. (Location 2636)

59. It is unlikely, however, that any workers affected could ever piece things together. First, there is nothing in writing. Second, Tucker feels sure that everyone involved would, if it became necessary, simply deny knowledge and claim that the process was altered solely for production reasons. 4. Finally, he says: The basic rule is that you hope that these kinds of things never occur. Nobody wants to hurt people. Nobody would ever consciously plan to do something that would endanger people. But when things happen, well, you cover for yourself and your company. (Location 2948)

60. Discreet suggestions, hints, and coded messages take the place of command; this, of course, places a premium on subordinates’ abilities to read correctly their bosses’ vaguely articulated or completely unstated wishes. One cannot even criticize one’s subordinates to one’s own superior without risking a negative evaluation of one’s own managerial judgment. (Location 3009)

61 (Chart). Phrase / Probable Intended Meaning

*Exceptionally well qualified / Has committed no major blunders to date

Tactful in dealing with superiors / Knows when to keep his mouth shut

Quick thinking / Offers plausible excuses

Meticulous attention to detail / A nitpicker

Slightly below average / Stupid

Unusually loyal / Wanted by no one else

Indifferent to instruction / Knows more than one’s superior

Strong adherence to principles / Stubborn

Requires work-value attitudinal readjustment / Lazy and hardheaded (Location 3018)

62. For the most part, euphemistic language is not used with the intent to deceive. Managers past a certain point, as suggested earlier, are assumed to be “maze-bright” and able to “read between the lines” of a conversation or a memorandum and to distinguish accurately suggestions from directives, inquiries from investigations, and bluffs from threats. Managers who are “maze-dense,” like the manager at Weft Corporation who, though told somewhat indirectly that he was fired, did not realize his fate until the following day, might consider the oblique, elliptical quality of managerial language to skirt deceit. However, most often when managers use euphemistic language with each other (and it is important to remember that in private among trusted others their language can be very direct, colorful, and indeed earthy), its principal purpose is to communicate certain meanings within specific contexts with

the implicit understanding that should the context change, a new, more appropriate meaning can be attached to the language already used. In this sense, the corporation is a place where people are not held to what they say because it is generally understood that their word is always provisional. (Location 3034)

63. The rule of thumb here seems to be that the more troublesome a problem, the more desiccated and vague the public language describing it should be. Of course, when a troublesome problem bursts into public controversy, euphemism becomes a crucial tool of those managers who have to face the public in some forum. The task here is to defuse public criticism and sometimes outrage with abstract unemotional characterizations of issues. (Location 3044)

F. Short Term Thinking

64. The ideal situation, of course, is to end up in a position where one can fire one's successors for one's own previous mistakes. (Location 2102)

65. Moreover, the only real threat to corporations on environmental issues was in the courts, which, however, judge past actions, not present practices. By the time the courts get to cases generated by contemporary practices, typically in fifteen years, those executives presently in charge will have moved on, leaving any problems their policies might create to others. (Location 707)

66. These choices also reflect judgments about whether a corporation can offer the hard-charging MBAs from the top-ranked schools enough quick variety to retain them long enough to justify their inflated salaries. (Location 926)

67. Similarly, accounting systems that place a premium on bare-bones inventory reflect the same pressure for short-run profit maximization. For instance, at Covenant Corporation the story is told about a plant that produced a useful by-product at no extra cost. One simply had to store it until it was needed for other internal operations. Covenant, however, works with an accounting system that considers by-products as inventory; moreover, inventory counts against one at the end of a fiscal year. In order to cut costs, managers decided to throw out the by-product at the end of a financial cycle. But a sudden shortage of the material trebled its cost two months later. To service their own operations, managers had to go hat in hand to their competitors to buy the material at the premium prices. (Location 1837)

68. This sets the stage for financial sharpshooters who, in takeover strategies, buy large chunks of a company's stock at devalued prices only to be "greenmailed" (persuaded with financial inducements) by the target company's management into surrendering these blocks of holdings at premium prices. In such unsettled times, where virtually any large corporation could become a takeover target, managers feel that they have to keep their companies' stock properly valued. As it happens, the markets honor only short-term gains. (Location 1854)

69. Instead, one focuses attention on important problems of the moment that must be solved. Since these are always plentiful, they justify postponing less pressing concerns. Of course, managers know at one level of their consciousness that today's minor issues can quickly become tomorrow's major crises, but the pressure for annual, quarterly, monthly, daily, and even hourly "results," that is, measurable progress plausibly attributed to one's own efforts, crowds out reflection about the future. An upper-middle manager at Alchemy Inc. recalls, for instance, his days as a plant

manager when his boss at company headquarters phoned him every three hours to see how many tons of soda ash had been produced in the interval. (Location 1869)

70. This goes to the heart of the problem. Managers think in the short run because they are evaluated by both their superiors and peers on their short-term results. Those who are not seen to be producing requisite short-run gains come to be thought of as embarrassing liabilities. Of course, past work gets downgraded in such a process. The old saw, still heard frequently today, “I know what you did for me yesterday, but what have you done for me lately?” is more than a tired garment district salesman’s joke. It accurately reflects the widespread amnesia among managers about others’ past accomplishments, however notable, and points to the probationary crucibles at the core of managerial life. Managers feel that if they do not survive the short run, the long run hardly matters, and one can only buy time for the future by attending to short-term goals. As one manager says: “Our horizon is today’s lunch.” (Location 1875)

71. Now you see this at work with mistakes. You can make mistakes in the work you do and not suffer any consequences. For instance, I could negotiate a contract that might have a phrase that would trigger considerable harm to the company in the event of the occurrence of some set of circumstances. The chances are that no one would ever know. But if something did happen and the company got into trouble, and I had moved on from that job to another, it would never be traced to me. The problem would be that of the guy who presently has responsibility. And it would be his headache. There’s no tracking system in the corporation. Some managers argue that outrunning mistakes is the real meaning of “being on the fast track,” the real key to managerial success. The same lawyer continues: In fact, one way of looking at success patterns in the corporation is that the people who are in high positions have never been in one place long enough for their problems to catch up with them. They outrun their mistakes. That’s why to be successful in a business organization, you have to move quickly. (Location 2013)

72. Both Covenant Corporation and Weft Corporation, for instance, place a great premium on a division’s or a subsidiary’s return on assets (ROA); managers who can successfully squeeze assets are first in line, for instance, for the handsome rewards allotted through bonus programs. One good way for business managers to increase their ROA is to reduce assets while maintaining sales. Usually, managers will do everything they can to hold down expenditures in order to decrease the asset base at the end of a quarter or especially at the end of the fiscal year. The most common way of doing this is by deferring capital expenditures, everything from maintenance to innovative investments, as long as possible. Done over a short period, this is called “starving a plant”; done over a longer period, it is called “milking a plant.” (Location 2034)

73. A plant that is not well maintained will fail in the short term, so you have to spend money there; a plant that has poorly trained people will fail today, so you have to spend money there. But you can still milk (Location 2045)

74. We’re judged on the short-term because everybody changes their jobs so frequently. (Location 2042)

75. My favorite things are not to replace my stores inventory and that shows up as direct profit on your balance sheet; not replace people who retire, and stretch everybody else out; cut down on overtime; cut working inventories to the bone. [You can also] lower the quality standards; you can get away with this in the short term

because people will accept that for awhile, though in the long term people will stop buying from you. (Location 2050)

76. At the very top of organizations, one does not so much continue to outrun mistakes as tough them out with sheer brazenness. In such ways, bureaucracies may be thought of, in C. Wright Mills's phrase, as vast systems of organized irresponsibility. (Location 2123)

77. When a weaver is unable to repair a stop, she shuts down the loom and flags a "fixer," the most skilled and highest-ranking worker on the shop floor, who does the repair or basic maintenance necessary to get the loom working again. However, since weavers are paid by the piece and can make no gain from a loom out of service for repair or maintenance during their own shifts, weavers will tend to remedy stops in any way they can in order to keep their cloth production high, leaving the maintenance and repair of the machinery and the economic cost involved to another weaver on another shift. Supervisors and managers who are evaluated by a plant's overall weaving efficiency are thus forced to monitor the number of loom stops constantly in order to make sure that looms badly in need of repair or maintenance get proper attention. Whenever structural inducements place premiums on immediate personal gains, especially when mistakes are not penalized, there seems to be a sharp decline in the likelihood of men and women sacrificing their own interests for others, for their organizations, or least of all for the common weal.2 (Location 2135)

78. It was said that Noll had milked and milked thoroughly every plant he ever supervised. One day, a story goes, he was accused of this in a public meeting by a vice-president who was then his superior. Noll is said to have responded with great boldness: "[Joe], how can you sit there and say that to me? How in the hell do you think you got to where you are and how do you think you stay there?" (Location 2153)

79. If a guy keeps moving, he can say, "Look, I ran this plant better than my predecessors." And people have to concede that. A lot of people do that. Then you get the guy who takes his place and tries to run things right and he has to spend a lot of money. And people look at the guy who was there before and they say: "Well, old [Noll] ran the plant well and he didn't have to spend any money like you're claiming you do." I don't think there is anything wrong with milking a plant. As long as you know you're milking (Location 2223)

80. No, definitely not. Would any sane, rational man spend $15 million for a 2 percent return? … Now it does improve the dust levels, but it was that if we don't invest the money now, we would be in a desperate [competitive] position fifteen years from now. Our demonstrated cash flow situation was such that eventually we would have had even tougher decisions to make. (Location 3579)

81. Executives also admit, somewhat ruefully and only when their office doors are closed, that OSHA's regulation on cotton dust has been the main factor in forcing the pace of technological innovation in a centuries-old, hidebound, and somewhat stagnant industry. It has also been a major factor in forcing executives to think in the long run rather than continually succumbing to short-term pressures. This is one of the reasons why the shrewdest among them only feigned elation at the attempts by Reagan's OSHA appointees to remove the cotton dust regulation from the purview of the Supreme Court. If such a move were successful, it could only encourage the traditionally reactionary elements of the textile industry who refuse to recognize on principle that government regulation, within reason, can be the businessman's best friend. (Location 3593)

82. S&Ls could now open transaction accounts and make commercial real estate loans, regardless of geography, up to 40 percent of their assets. This prompted many S&L executives to sell their long-term, fixed-rate, low- or no-profit mortgages to Wall Street firms for as little as 60 percent of their value in order to obtain ready funds for much more lucrative, though much riskier, investments. (Location 4649)

83. Organized irresponsibility has migrated to the nooks and crannies of our society. One sees presidents of universities and colleges who preside over the destruction of their institutions' endowments, not to mention the intellectual integrity of curricula, and then outrun their mistakes and move on to higher academic and even top political posts. One sees intellectuals who are completely committed to the destruction of the Western tradition that nurtures them and who help create the profound social and cultural centrifugality that marks American society. One sees big-city mayors who cede control of the streets to drug dealers and other thugs for fear of being labeled nonprogressive by various advocacy groups and the press. One sees long-time political insiders in Washington who "screw up and move up," an aphorism that aptly captures the ethos of government bureaucracies. And doublethink and doublespeak now permeate public discourse, bewildering even intelligent, thoughtful citizens. (Location 4990)

G. Social Politics

84. Managers' cryptic aphorism, "Well, you never know…," repeated often and regularly, captures the sense of uncertainty created by the constant potential for social reversal. Managers know too, and take for granted, that the personnel changes brought about by upheavals are to a great extent arbitrary and depend more than anything else on one's social relationships with key individuals and with groups of managers. (Location 761)

85. "Every big organization is set up for the benefit of those who control it; the boss gets what he wants." (Location 827)

86. In any event, just as managers must continually please their boss, their boss's boss, their patrons, their president, and their CEO, so must they prove themselves again and again to each other. Work becomes an endless round of what might be called probationary crucibles. Together with the uncertainty and sense of contingency that mark managerial work, this constant state of probation produces a profound anxiety in managers, perhaps the key experience of managerial work. (Location 910)

87. See, once you are at a certain level of experience, the difference between a vice-president, an executive vice-president, and a general manager is negligible. It has relatively little to do with ability as such. People are all good at that level. They wouldn't be there without that ability. So it has little to do with ability or with business experience and so on. All have similar levels of ability, drive, competence, and so on. What happens is that people perceive in others what they like—operating styles, lifestyles, personalities, ability to get along. Now these are all very subjective judgments. And what happens is that if a person in authority sees someone else's guy as less competent than his own guy, well, he'll always perceive him that way. And he'll always pick—as a result—his own guy when the chance to do so comes up. (Location 1013) [Also see Nothing Matters]

88. Striking, distinctive characteristics of any sort, in fact, are dangerous in the corporate world. One of the most damaging things, for instance, that can be said about a manager is that he is brilliant. This almost invariably signals a judgment that

the person has publicly asserted his intelligence and is perceived as a threat to others. What good is a wizard who makes his colleagues and his customers uncomfortable? (Location 1173)

89. Equally damaging is the judgment that a person cannot get along with others—he is “too pushy,” that is, he exhibits too much “persistence in getting to the right answers,” is “always asking why,” and does not know “when to back off.” Or he is “too abrasive,” or “too opinionated,” unable “to bend with the group.” Or he is a “wildman” or a “maverick,” that is, someone who is “outspoken.” Or he may be too aloof, too distant, “too professional.” (Location 1176)

90. The knowledgeable practitioners of corporate politics, whether patrons or leaders of cliques and networks, value nothing more highly than at least the appearance of unanimity of opinion among their clients and allies, especially during times of turmoil. (Location 1197)

91. Managers know that patrons and powerful allies protect those already selected as rising stars from the negative judgments of others; and only the foolhardy point out even egregious errors of those in power or those destined for it. (Location 1463)

92. Managers know that to be weak in a world that extols strength and power is to invite abuse. (Location 1536)

93. Such social distancing has two purposes: it undermines in advance or lays the groundwork for refusal of any claims that a person considered a failure might make on another, and it forestalls the possibility of being linked with that person in others’ cognitive maps. This becomes particularly important when there has been a known past association between oneself and one thought to have failed in some way. (Location 1542)

94. In fact, when it is even suspected that a person might be headed for trouble, anticipatory avoidance is the rule. (Location 1550)

95. Being a “fall guy,” that is, “taking the rap” or “taking the heat” for others’ decisions or mistakes is probably the most common kind of blame in big organizations. (Location 1904)

96. You have to remember that you only get to explain things away once. When things get screwed up, you get one chance. That’s why it’s important for everybody to be in bed with everybody else. And if they don’t like you from the start, you don’t have a chance. Because when things go wrong, what people do is sit down and say—without saying it in so many words—look, our jobs are on the line. Let’s make sure that it’s not us who gets nailed. (Location 1999)

97. Wilson and his Site Operations staff were particularly concerned because when the polar crane was finally used, it would be Site Operations who used it. They knew that, if they were in charge when the button was pushed, they could be blamed for whatever might go wrong. If the polar crane failed, it would be seen as the fault of Site Operations. (Location 2557)

98. high-ranking staff official at Covenant explains: Anxiety is endemic to anyone who works in a corporation. By the time you get to be middle management, it’s difficult to make friends because the normal requirement for friendship—that is, loyalty—doesn’t fit in this context. You have to look out for number one more than anything else. Moreover, the prevailing view is that managers are big boys and girls, well-paid, and

should be able to take care of themselves. Besides, one person’s failure represents another person’s opportunity. (Location 1554)

99. of a young manager who, while at a company conference, went out for his weight-controlling 5:30 A.M. jog only to meet a vice-president similarly engaged, a powerful executive who now cheers the younger man’s work and presentations and introduces him to other influential senior managers; (Location 1654)

100. scattering ducks already set in a row. Besides, one can only criticize something when one has the resources to solve it in a clear and decisive way. Otherwise, one should keep one’s skepticism to oneself and get “on board.” (Location 2629)

101. Other managers and managerial cliques are always on the lookout for others’ mistakes or for actions that can be construed as mistakes and will pounce on anyone foolish enough to admit them. Even if others restrain an immediate attack, the knowledge of someone’s mistakes is ammunition for the future. Many managers “lay in the weeds, with rocks, and wait.” One who exposes a colleague’s errors in such a context and makes him vulnerable to others evinces, of course, only a fundamental untrustworthiness, unless one’s colleague has first betrayed oneself or others in some way— (Location 2858)

102. There is a premium in the higher circles of management on seeming fresh, dynamic, innovative, and up-to-date. In their social minglings and shoptalk with one another, particularly with their opposite numbers in other large companies, say, at the Business Roundtable, at high-level conferences at prestigious business schools, at summer galas in the Hamptons, or at the Super Bowl, the biggest business extravaganza of all, executives need to seem abreast of the latest trends in managerial know-how. No one wants to appear stodgy before one’s peers nor to have one’s firm defined in managerial networks, and perhaps thence to Wall Street, as “slow on the uptake.” Executives trade ideas and schemes and judge the efficacy of consultant programs not by any detached critical standards but by what is socially acceptable, desirable, and, perhaps most important, current in their circles. (Location 3155)

103. So they attend the sessions and with a seemingly dutiful eagerness learn literally to repeat the requisite formulas under the watchful eyes of senior managers. Senior managers do not themselves necessarily believe in such programs. In one seminar that I attended, the senior manager in charge startled a room of juniors by saying: Fellows, why aren’t any of you asking about the total lack of correspondence between what we’re preaching here and the way we run our company? But such outspokenness is rare. (Location 3209)

104. They admit, however, that the marvelously high fees that consultants command (in 1986 as high as $2,000 a day in New York City) enhance their legitimacy and encourage managers to lend credence to their schemes. (Location 3216)

105. A choice between securing one’s own success by jumping on and off the bandwagon of the moment, or sacrificing oneself for the long-run good of a corporation by diverting resources and really seeing a program through is, for most managers, no choice at all. Ambitious managers see self-sacrificing loyalty to a company as foolhardy. Moreover, middle and upper-middle level managers upon whom requests for self-sacrifice for the good of the organization are most likely to fall do not see top executives sacrificing themselves for the common good. For example, just after the CEO of Covenant Corporation announced one of his many purges,

legitimated by a “comprehensive assessment of the hard choices facing us” by a major consulting firm, he purchased a new Sabre jet for executives and a new 31-foot company limousine for his own use at $1,000 a foot. He then flew the entire board of directors to Europe on the Concorde for a regular meeting to review, it was said, his most recent cost-cutting strategies. As other managers see it, bureaucratic hierarchy gives top bosses the license to act in their own interests and to pursue with impunity the arts of contradiction. (Location 3220)

H. The I in Team

106. The main dimensions of team play are as follows. 1. One must appear to be interchangeable with other managers near one’s level. Corporations discourage narrow specialization more strongly as one goes higher. They also discourage the expression of moral or political qualms. One might object, for example, to working with chemicals used in nuclear power, or working on weapons systems, and most corporations today would honor such objections. Publicly stating them, however, would end any realistic aspirations for higher posts because one’s usefulness to the organization depends on versatility. As one manager in Alchemy Inc. comments: “Well, we’d go along with his request but we’d always wonder about the guy. And in the back of our minds, we’d be thinking that he’ll soon object to working in the soda ash division because he doesn’t like glass.” Strong convictions of any sort are suspect. One manager says: If you meet a guy who hates red-haired persons, well, you’re going to wonder about whether that person has other weird perceptions as well. You’ve got to have a degree of interchangeability in business. To me, a person can have any beliefs they want, as long as they leave them at home. Similarly, one’s spouse’s public viewpoints or activities could reduce others’ perceptions of a manager’s versatility or indeed ability. In reference to another manager whose wife was known to be active in environmental action groups, lobbying in fact for legislation on chemical waste disposal, one Alchemy manager says: “If a guy can’t even manage his own wife, how can he be expected to manage other people?” (Location 1137)

107. Team play also means, as one manager in the chemical company puts it, “aligning oneself with the dominant ideology of the moment” or, as another says, “bowing to whichever god currently holds sway.” Such ideologies or gods may be thought of as official definitions of reality. (Location 1188)

108. You can indict a person by saying that he’s not a team player. That doesn’t mean he won’t follow directions. It’s because he voices an objection, because he argues with you before doing something, especially if he’s right. That’s when we really get mad—when the other guy is right. If he’s wrong, we can be condescending and adopt the “you poor stupid bastard” tone.... (Location 1220)

109. A team player is a manager who does not “force his boss to go to the whip,” but, rather, amiably chooses the direction his boss points out. Managers who choose otherwise or who evince stubbornness are said to “have made a decision,” a phrase almost always used to describe a choice that will shorten a career. (Location 1229)

110. Team players display a happy, upbeat, can-do approach to their work and to the organization. (Location 1252)

I. Avoiding Decision Making

111. You know that old saying: “Success has many parents; failure is an orphan”? Well, that describes decision making. A lot of people don’t want to make a commitment, at least publicly. This is a widespread problem. They can’t make judgments. They stand around and wait for everybody else’s reaction. Let me tell you a story which perfectly illustrates this. There was a [museum] collection coming, the [Arctic] collection, and there was a great deal of interest among designers in [Arctic] things. My own feeling was that it wouldn’t sell but I also recognized that everybody wanted to do it. But in this case, [our] design department was spared the trouble. There was an independent designer who had access to our president and he showed him a collection of [Arctic] designs. There were two things wrong: (1) it was too early because the collection hadn’t hit town yet; (2) more important, the designs themselves were horrible. Anyway, [the collection] was shown in a room with everything spread out on a large table. I was called down to this room which was crowded with about nine people from the company who had seen the designs. I looked at this display and instantly hated them. I was asked what I thought but before I could open my mouth, people were jumping up and down clapping the designer on the back and so on. They had already decided to do it because the president had loved it. Of course, the whole affair was a total failure. The point is that in making decisions, people look up and look around. They rely on others, not because of inexperience, but because of fear of failure. They look up and look to others before they take any plunges. (Location 1713)

112. But all these things have no relationship to the way they actually manage or make decisions. The basic principles of decision making in this organization and probably any organization are: (1) avoid making any decision if at all possible; (2) if a decision has to be made, involve as many people as you can so that, if things go south, you’re able to point in as many directions as possible. (Location 1731)

113. A middle-aged, upper-middle level manager at Alchemy Inc. says: You know, there is this huge computerized inventory of skills which people update each year; it’s called a skills inventory. … But all the computerized lists in the world don’t amount to much in the corporation. What matters is a bunch of guys sitting informally in a room and deciding who should get jobs and who shouldn’t. The real job decisions are made on that basis. And circumstances determine your fate. (Location 1634)

114. Making a decision, or standing by a decision once made, exposes carefully nurtured images of competence and know-how to the judgments of others, particularly of one’s superiors. As a result, many managers become extremely adept at sidestepping decisions altogether and shrugging off responsibility, all the while projecting an air of command, authority, and decisiveness, leaving those who actually do decide to carry the ball alone in the open field. (Location 1774)

115. This explains why the chemical company managers kept putting off a decision about major reinvestment. After the battery collapsed in 1979, however, the decision facing them was simple and posed little risk. The corporation had to meet its legal obligations; also, it had either to repair the battery the way the EPA demanded or shut down the plant and lose several hundred million dollars. Since there were no real choices, everyone could agree on a course of action because everyone could appeal to inevitability. This is the nub of managerial decision making. As one manager says: Decisions are made only when they are inevitable. To make a decision ahead of the time it has to be made risks political catastrophe. People can always interpret the decision as an unwise one even if it seems to be correct on other grounds. (Location 1886)

116. Despite their fresh appearances, certain themes recur constantly in the programs offered by consultants. Perhaps the most common are how to sharpen decision making, how to restructure organizations for greater efficiency, how to improve productivity, how to recognize trouble spots in an organization, how to communicate effectively, how to humanize the workplace, and how to raise morale. (Location 3164)

J. This is Your Life

117. At the managerial and professional levels, the road between work and life is usually open because it is difficult to refuse to use one's influence, patronage, or power on behalf of another regular member of one's social coterie. It therefore becomes important to choose one's social colleagues with some care and, of course, know how to drop them should they fall out of organizational favor. (Location 884)

118. For some managers, the drive for success is a quest for the generous financial rewards that high corporate position brings. For others, success means the freedom to define one's work role with some latitude, to "get out from under the thumb of others." For still others, it means the chance to gain power and to exert one's will, to "call the shots," to "do it my way," or to know the curiously exhilarating pleasure of controlling other people's fates. For still others, the quest for success expresses a deep hunger for the recognition and accolades of one's peers. (Location 955)

119. More generally, those who accept immobility are unwilling to sacrifice family life or free-time activities to put in the extraordinarily long hours at the office required in the upper circles of their corporations. Or they have made a realistic assessment of the age structure, career paths, and power relationships above them and conclude that there is no longer real opportunity for them. They may see that there is an irreparable mismatch between their own personal styles and the kinds of social skills being cultivated in well-entrenched higher circles. In many cases, they decide that they do not wish to put up with the great stress of higher management work that they have witnessed. (Location 962)

120. Higher-level managers in all the corporations I studied commonly spend twelve to fourteen hours a day at the office. (Location 1156)

121. In a world where appearances—in the broadest sense—mean everything, the wise and ambitious manager learns to cultivate assiduously the proper, prescribed modes of appearing. He dispassionately takes stock of himself, treating himself as an object, as a commodity. He analyzes his strengths and weaknesses and decides what he needs to change in order to survive and flourish in his organization. And then he systematically undertakes a program to reconstruct his image, his publicly avowed attitudes or ideas, or whatever else in his self-presentation that might need adjustment. (Location 1339)

122. This [lack of control] doesn't mean you don't work hard; at least in my case, that's my answer. I have to believe I can influence events. That way, I feel good about myself even if my boss doesn't. (Location 1606)

123. Sometimes I'll wake up in the middle of the night thinking about the plant. And if I can't get back to sleep, I'll slip out of bed and walk over to the plant and just walk around the machinery and talk to the guys. I love the smell of the oil and the grease and the sound of the machines. For me, that's what life is all about. (Location 4535)

K. Immoral Mazes

124. In analyzing incidence data from a representative plant, and by extrapolating to the rest of the firm’s mills, White discovered that 12 percent of all greige mill workers had already suffered hearing loss severe enough to be immediately compensable under state law for as much as $3.5 to $5.7 million. This, however, was only the thunder before a summer storm. Another 63 percent of greige mill workers had already suffered substantial, though not yet compensable, damage that could only worsen the longer they stayed in the industry. (Location 2265)

125. Moreover, by insisting on his own personal moral purity, his feeling that if he did not expose things he himself would be drawn into a web of corruption, he was, they feel, being disingenuous; no one reaches his level of a hierarchy without being tainted. (Location 2450)

126. A moral judgment based on a professional ethic makes little sense in a world where the etiquette of authority relationships and the necessity for protecting and covering for one’s boss, one’s network, and oneself supercede all other considerations and where nonaccountability for action is the norm. (Location 2474)

127. On the polar crane issue, the top official of the NRC later characterized Wilson’s and his engineers’ concerns as stemming from a philosophy that emphasized procedural matters rather than a focus on final goals. This characterization was echoed by GPUN management who stressed that what was at issue in the polar crane dispute was not procedures but results; at a certain point, they said, decisions had to be made to resolve technical disputes and work had to proceed toward what everyone acknowledged to be a worthwhile goal, that is, the cleanup of TMI-2. (Location 2566)

128. Once, when he wrote a memo to the GPUN deputy director about radioactively contaminated sewage being trucked out of the plant and disposed of illegally, his boss replied that he did not need such a memo from Wilson. It was, his boss said, not constructive and wasted his own and Wilson’s time. (Location 2575)

129. my view, this is the nub of the moral ethos of bureaucracy. Managers see this issue as a “trade-off” between principle and expediency. They usually pose the trade-off as a question: Where do you draw the line? (Location 2652)

130. It is said, in fact, that one Weft manager who was on the take for years from customers was fired within a half hour of the discovery of his “inexplicably stupid” acceptance and deposit of a check from his benefactors instead of his normal cash rake-offs. With such authoritative encouragement, managers internalize these norms into their own personal codes of honor; they speak privately of the importance of not being known as men or women “who can be had.” (Location 2658)

131. It gets hard. Now, suppose that the ozone depletion theory were correct and you knew that these specific fifty people were going to get skin cancer because you produced chlorofluorocarbons [CFCs]. Well, there would be no question. You would just stop production. But suppose that you didn’t know the fifty people and it wasn’t at all clear that CFCs were at fault, or entirely at fault. What do you do then? (Location 2831)

132. Suppose that you had a candy bar factory and you were touring the plant and you saw with your own eyes a worker slip a razor blade into a bar. And before you could stop the machine, there were a thousand bars more made and the one with the

razor blade was mixed up. Well, there’s no question that you would get rid of the thousand candy bars. But what if it were a million bars? Well, I don’t know what I’d do. (Location 2843)

133. Moreover, he tries to treat his subordinates forthrightly, firmly believing that one’s word is an important measure of a person. In a world, however, where actions are separated from consequences, where knowledge is fragmented and secreted, where private agreements are the only real way to fashion trust in the midst of ongoing competition and conflict, where relationships with trusted colleagues constitute one’s only real means both of defense and opportunity, and where, one knows, even coincidental association with a disaster can haunt one’s career years later, keeping silent and covering for oneself and for one’s fellows become not only possible but prudent, indeed virtuous, courses of action. (Location 2969)

134. Moreover, it is a byword in the chemical industry at least that it is precisely in those technological areas where accidents have seldom occurred that the largest potential catastrophes loom; the very lack of practice in responding quickly to untoward incidents can precipitate uncontrollable events. (Location 3376)

135. Publicly, of course, Weft Corporation, as do many other firms, claims that the money was spent entirely to eliminate dust, evidence of its corporate good citizenship. Privately, executives admit that without the productive return, they would not have—indeed, given the constraints under which they operate—could not have spent the money. (Location 3585)

136. The ethic that emerges in bureaucratic contexts contrasts sharply in many respects with the original Protestant ethic. The Protestant ethic was a social construction of reality of a self-confident and independent propertied social class. It was an ideology that extolled the virtues of accumulating and reinvesting wealth in a society organized around property and that accepted the stewardship responsibilities entailed by property. It was an ideology where a person’s word was his bond and where the integrity of the handshake was crucial to the maintenance of good business relationships. (Location 4295)

137. At the core of the middle class’s righteous, some would say smug, faith in itself, of its inexhaustible drive, of its unremitting pragmatism, was the conviction that hard work necessarily had its just rewards here and now as a token of divine favor in the hereafter. This conviction was also the bedrock of a profound guilt mechanism that impelled one to fulfill personal and social obligations; failure to do so, like a failure to work hard, was thought to be a sin against both God and self. (Location 4304)

138. Bureaucracy breaks apart the ownership of property from its control, social independence from occupation, substance from appearances, action from responsibility, obligation from guilt, language from meaning, and notions of truth from reality. Most important, and at the bottom of all of these fractures, it breaks apart the older connection between the meaning of work and salvation. In the bureaucratic world, one’s success, one’s sign of election, no longer depends on an inscrutable God, but on the capriciousness of one’s superiors and the market; and one achieves economic salvation to the extent that one pleases and submits to new gods, that is, one’s bosses and the exigencies of an impersonal market. (Location 4307)

L. Truth and Public Relations

139. Truth? What is truth? I don’t know anyone in this business who talks about “truth.” (Location 4137)

140. After all, as one executive says, “We lie all the time, but if everyone knows that we’re lying, is a lie really a lie?” (Location 2693)

141. If a retailer returns with goods and says, “Look, I can’t sell $50,000 of this stuff. You have to take it back or it’s going to break me,” may one say, “What’s that? I didn’t hear you,” hoping for the sake of future business that the retailer will say, “Oh, look, you were late delivering and I can’t use it now.” (Location 2700)

142. The consultant who perceives such discrepancies has to devise his own strategies for handling them. Some of these include: rejecting the assignment altogether; accepting the problem as defined and confining oneself to it for the sake of future contracts even though one knows that any action will be inefficacious; or accepting the assignment but trying to persuade the client to address the underlying social and political issues, that is, redefining the problem. (Location 3234)

143. For instance, a highly placed staff member whose work requires him to interact daily with the top figures of his company, says: I get faked out all the time, and I’m part of the system. I come from a very different culture. Where I come from, if you give someone your word, no one ever questions it. It’s the old hard-work-will-lead-to-success ideology. Small community, Protestant, agrarian, small business, merchant-type values. I’m disadvantaged in a system like this. He goes on to characterize the system more fully and what it takes to succeed within it: It’s the ability to play this system that determines whether you will rise. … And part of the adeptness [required] is determined by how much it bothers people. One thing you have to be able to do is to play the game, but you can’t be disturbed by the game. What’s the game? It’s bringing troops home from Vietnam and declaring peace with honor. It’s saying one thing and meaning another. It’s characterizing the reality of a situation with any description that is necessary to make that situation more palatable to some group that matters. It means that you have to come up with a culturally accepted verbalization to explain why you are not doing what you are doing…. [Or] you say that we had to do what we did because it was inevitable; or because the guys at the [regulatory] agencies were dumb; [you] say we won when we really lost; [you] say we saved money when we squandered it; [you] say something’s safe when it’s potentially or actually dangerous…. Everyone knows that it’s bullshit, but it’s accepted. This is the game. (Location 3246)

144. Such adeptness at inconsistency, without moral uneasiness, is essential for executive success. Done over a period of time, in fact, it seems to become a taken for granted habit of mind. As one executive at Covenant Corporation says: Now some people don’t understand this…. But as you move up the ladder, you don’t have people who don’t understand. And the people up high don’t necessarily do it consciously. They are able to speak out of both sides of their mouth without missing a step. It means being able to say, as a very high-ranking official of Weft Corporation said to me without batting an eye, that the industry has never caused the slightest problem in any worker’s breathing capacity. It means, in Covenant Corporation, propagating an elaborate hazard/benefit calculus for approval of dangerous chemicals while internally conceptualizing “hazards” as business risks. It means publicly extolling the carefulness of testing procedures on toxic chemicals while privately ridiculing animal tests as inapplicable to humans. (Location 3610)

145. Finally, those truly adept at inconsistency can also interpret with some accuracy the inconsistent machinations of their colleagues and adversaries. This is not a mean skill. At the very beginning of my fieldwork, the top lawyer of a large corporation was discussing an issue that I had raised when he said: Now, I'm going to be completely honest with you about this. He paused for a moment and then said: By the way, in the corporate world, whenever anybody says to you: "I'm going to be completely honest with you about this," you should immediately know that a curveball is on the way. But, of course, that doesn't apply to what I'm about to tell you. (Location 3631)

146. Since the success of large commercial bureaucracies depends to a great extent on the goodwill of the consuming public, ambitious managers recognize that great organizational premiums are placed on the ability to explain expedient action convincingly. Public opinion, of course, constitutes one of the only effective checks on the bureaucratic impulse to translate all moral issues into practical concerns. (Location 3643)

147. 1925, Walter Lippmann critically analyzed a public's ability to appreciate the intricacies of any issue or indeed to keep its attention on anything but crises. He adds: [S]ince [a public] acts by aligning itself, it personalizes whatever it considers, and is interested only when events have been melodramatized as a conflict. The public will arrive in the middle of the third act and will leave before the last curtain, having stayed just long enough perhaps to decide who is the hero and who is the villain of the piece. (Location 3765)

148. read in The Wall Street Journal or The New York Times accounts of specific events or larger trends attributed to "informed sources," the code name for public relations men and women; (Location 3828)

149. There are, of course, good clients and bad clients. A good client keeps his public relations specialists adequately informed, provides "feedback" and "constructive criticism," and recognizes that public relations depends on time-billing rather than on product billing and provides enough funds to do the job properly. A good client "takes stock of himself," that is, dispassionately objectifies his company's products, his company's organizational structure and personnel, and, say, in the case of a top corporate official, himself, to make them all more readily manipulable and therefore marketable commodities. Above all, a good client is flexible and therefore able rapidly to shift ground and actual policies as well to meet new needs or new pressures. A bad client, by contrast, either does not understand or chooses to ignore the peculiarly indirect approach of public relations and wants immediately observable results in terms of press clippings or TV time; uses a public relations program as a vehicle for self-aggrandizement within his own corporation, placing the public relations specialist in danger of "getting caught in a pissing contest between executives"; demands the release of press statements that are only marginally "newsworthy" and then blames the public relations specialist when nothing at all gets printed or aired; conceals crucial aspects of his story from his public relations advisers and refuses them adequate access to his staff and facilities to get the full story; comes to the public relations agency with trumped-up data and fake photographs—as happened, for instance, at Images Inc. when a client falsely claimed the efficacy of a product to remove tar from despoiled beaches—and ends up declaring bankruptcy, leaving the public relations agency liable for a multimillion-dollar lawsuit; wants the public relations agency, as in the case of some single-party foreign governments, to put a good public face on practices like the "persuasion" and "elimination" of opponents; expects the public relations agency to be the "bag man" to pay off government officials or newspaper editors; and, perhaps especially, insists on an indefensible or

totally unbelievable version of reality or expects the public relations specialist to tell outright lies. (Location 3843)

150. As it happens, accounts in the public relations world circulate continually and only sometimes because of actual or alleged dissatisfaction with public relations service. The pattern of continual mergers, upheavals, and power struggles in corporations, that I have argued earlier is at the core of American corporate life, directly affects public relations agencies. When a new CEO or divisional president assumes power in a corporation, he will quite often change public relations and advertising agencies as part of a larger strategy of shedding the past or to assert his own “new vision” of the future. When this occurs, almost invariably, he will move his account to an agency whose leaders he knows and with whom he feels comfortable. This continual circulation of accounts means that agency personnel are constantly searching for new accounts and constantly devising ways to hold present clients, despite the knowledge of the inevitability of their eventual departure. (Location 3868)

151. In the world of public relations, there is no such thing as a notion of truth; there are only stories, perspectives, or opinions. One works, of course, with “facts,” that is, selected empirically verifiable statements about the world. But as long as a story is factual, it does not matter if it is “true.” One can feel free to arrange these facts in a variety of ways and to put any interpretation on them that suits a client’s objectives. Interpretations and judgments are always completely relative. The only canons binding this process of interpretation are those of credibility or, more exactly, of plausibility. If an interpretation of facts, a story, is taken as plausible by a targeted audience, it is just as good as “true” in any philosophical sense, indeed better since it furthers the accomplishment of an immediate goal. (Location 3886)

152. From the standpoint of public relations, the journalistic ideology closely resembles the social outlook of most college seniors—a vague but pious middle-class liberalism, a mildly critical stance toward their fathers in particular and authorities in general; a maudlin championship of the poor and the underclass; and especially the doctrine of tolerance, open-mindedness, and balance. In fact, public relations people feel, the news media are also constructing reality. They are always looking for a “fresh” and exciting angle; they have an unerring instinct for the sentimental that expresses itself in a preference for “human interest” rather than substance; and they arrange facts in a way that purports to convey “truth,” but is in fact simply another story. In reality, news is entertainment. And, despite the public’s acceptance of journalistic ideologies, most of the public watch or read news not to be informed or to learn the “truth,” but precisely to be entertained. There is no intrinsic reason, therefore, why the constructions of reality by public relations specialists should be thought of as any different from those of any group in the business of telling stories to the public. (Location 3903)

153. The top official, in a written document prepared for a meeting to discuss the issue, argued that: [T]he final report was better than it would have been if it had not gone through [the] process of reexamination. It was stronger, it was more important, it was more constructive…. At the same time, there is no doubt that pressures were brought on us to come to the kind of conclusions we finally reached. At the meeting, he added, somewhat more pithily: “We tried to be honest but, believe me, it wasn’t easy.” (Location 4124)

154. The same top executive defines truth: I sometimes sit back and think that if we could make up a list of all the viewpoints of all our clients and somehow fit them

together, then that would be truth. That would be what we are as a firm. (Location 4130)

155. Agency practitioners not only defend multiple interests but face repeatedly the experience of even well-served clients switching agencies. Especially unsettling are the departures of clients who decide that, since they are as they have been portrayed, they have no future need to construct reality. The continual circulation of client accounts in agencies diminishes the possibilities of comforting, long-held allegiances to organizations, products, or causes. (Location 4192)

156. Alternatively, and by contrast, practitioners in both settings sometimes justify their efforts by appealing to a professional ethos that celebrates the exercise of technical skill separated from any emotional commitment to one's clients. A dignified version of this legitimation is the often repeated analogy between public relations practitioners and lawyers; both occupations, it is argued, fulfill important advocacy roles in a free society. Only the practice of the professional virtue of public relations, however hazardous to individual practitioners, can assure the continued diversity of opinion that marks our democracy. (Location 4195)

157. That's the reality of our society. There's no question that their story is being told because they have the money and power. I've got to recognize that I'm part of this society and just come to live with that. Our society is the way it is. It's run on money and power, it's that simple. Truth has nothing to do with it. So we just accept the world as it is and live with it. (Location 4206)

158. The agency literally creates these realities by matchmaking artistic talent and accomplishment with money. In the process, up-and-coming junior corporate executives get the kind of exposure to refined, sophisticated artistic and intellectual circles that will help prepare and polish them for higher posts. Artists, in turn, as long as their work is not too avant-garde, receive the benefits of a latter-day Medici-like patronage. Public relations people claim a double accomplishment—they help civilize businessmen, not least by inspiring in them a "passion for greatness" rather than self-interest as the important motive for patronage of the arts; and they provide the public with access to high culture. The same firm has also arranged corporate sponsorship of dance and musical performances, helped develop important educational programs such as one on infant nutrition, conducted some useful surveys such as one on the problems facing ethnic minorities, and done a lot of pro bono work for philanthropic, community, and public service organizations in the bargain. One tries, then, to move some clients in directions that seem socially desirable while at the same time playing with the magic lantern to serve their interests. Sometimes, too, the ability to accomplish any good at all in this world seems to depend on the willingness to serve even clients with no apparent redeeming features in order to seize capricious opportunities to channel other clients' resources into work deemed socially worthwhile. (Location 4214)

159. In short, bureaucracy creates for managers a Calvinist world without a Calvinist God, a world marked with the same profound anxiety that characterized the old Protestant ethic but one stripped of that ideology's comforting illusions. Bureaucracy poses for managers an intricate set of moral mazes that are paradigmatic of the quandaries of public life in our social order. Within this framework, the puzzle for many individual managers becomes: How does one act in such a world and maintain a sense of personal integrity? (Location 4351)

M. Under Siege

160. Part of the folklore of the modern corporation, in fact, consists of a catalog of stories about how big corporations are being victimized by the courts in what amounts to a radical transformation of tort law. Managers repeat variations of cases—for instance, of a man who purchased a thrice-owned punch machine, altered it, and after losing a finger, sued the original manufacturer for negligent construction; of a farmer who, despite clearly marked warning labels, coated his animal feed troughs with a toxic wood preservative, sold milk from contaminated cattle, and then when he himself was sued by irate parents whose children had drunk the milk, sued the chemical company who produced the preservative; or of a student research assistant inexplicably asked to carry dangerous acid in glass containers on a faulty elevator that lurched and caused the young woman to break a container, showering her with acid and permanently disfiguring her. Although she is said to have sued her university, her supervising professor, the elevator manufacturer, and the chemical company that provided the acid, managers had few doubts that the jury would “pin the tail on business” because of the “abiding conviction that corporations have vaults filled with gold bars called profits.” At some level, even managers who repeat such stories and thus reinforce their own shared sense of beleaguerment, recognize the profound and profoundly felt dependence of most people in our society on those large organizations that claim to control the scientific genie. (Location 3413)

161. was at this party the other night and I was sitting next to this older lady and she said: “My God, did you people see the paper tonight? There’s leakage from some chemical plant and it’s infecting the drinking water around here.” So I asked if I could see the paper and the article said that there was some seepage out of a pond that a chemical plant used for disposal but that the EPA was monitoring the situation as well as all 25 wells in the area. In the meantime, the company was remedying the situation. I pointed this out and she looked at me, her eyes narrowed, and she asked, “Who do you work for?” I said I worked for a large chemical firm and she burst out laughing and asked if I expected her to believe me. She laughed right in my face! I asked her what she was worried about. You have to drink the water for 25 years before anything would happen—I mean, she was already a grandmother—but that didn’t seem to help her much. (Location 3459)

162. My interviews are filled with stories of managers who claim to have been verbally assaulted not only by strangers at cocktail parties, but by their children’s teachers when they visit schools, and even by their children themselves at the breakfast table for being supposedly callous and insensitive to the social consequences of business activity. Perhaps more than anything else, managers are puzzled by such attacks, though they pounce quickly on any inconsistencies that they perceive in their opponents. For example, one manager whose firm produced a pesticide that became caught up in a widely publicized episode of mishandling and illegal disposal was attacked by his brother-in-law, a lifelong military man, for his very association with the company. The manager found it grimly ironic that “things have reached a point where a trained killer is berating me for producing something useful.” More generally, managers view the irrational fear of contamination evinced by those who espouse an ideology of bodily purity with some derision. One can, after all, referring to the same pesticide, “eat handfuls of the stuff with no lasting adverse effects.” (Location 3467)

163. Well, from 1957 through 1962, I was intimately involved with the manufacture of DDT. During that time, we doubled production and sold almost all of it to Africa and India. And I knew and went home knowing that I was saving more lives than any major hospital was capable of doing. I knew that I was saving thousands of lives by doing this. Then Rachel Carson’s Silent Spring came out and not only did I become a murderer of falcons and robins, but also one of the mass murderers of the world. I was

now doing evil things to the world. Then I went to a plant which was manufacturing [chlorofluorocarbons] and we increased production by 20–25 percent; a lot of it went into hair sprays. We also used vinyl chloride and found out that it was causing liver cancer. Then I found out that I was destroying the whole ozone layer of the earth and doing it for personal gain. Then I went into soda ash. Without it, there wouldn't be a window pane in the whole United States. But at the time, because of the Clean Air Act and the Clean Water Act, suddenly I became a polluter. Children learned in school that chemicals killed. And, of course, there was no question in the academic community that I was perceived as an evil person doing evil things. And that became true even in the corporation. Plants became a liability, rather than the source of wealth. The perception was: Wouldn't it be nice if we could just sell chemicals without producing them? So the profession of producing things became a low profession and the good people were those who were producing services. Manufacturing people became evil. I think this is one reason that marketing people became ascendant in the competition for advancement. Then I went into making sulfuric acid and that involved the whole issue of water pollution. He goes on to describe his reaction to all of this: I guess I'm more bemused than anything else. It's the same type of feeling that I would have if I were an MD who had been doing radical mastectomies and then someone says—hey, you didn't have to do that. It's a feeling of disappointment without a feeling of shame. I know that I have done some useful things. I know that the only source of money is taking something out of the ground and making something. That's where money comes from—making something out of nothing. Without that type of activity, civilization wouldn't exist. So, on something like DDT, I'll leave the judgment to Europe after World War II and all the people saved by the widespread use of it. (Location 3482)

164. Within this context of perceived harassment and shifting scientific and ideological winds, while always attending to the pressing and sometimes contradictory exigencies of business life, managers must address a multiplicity of audiences, some of whom are considered rivals, and some outright adversaries. These audiences are the internal corporate hierarchy with its intricate and shifting power cliques and competing managerial circles, key regulators, local and federal legislators, special publics that vary according to the issues, and the public at large, whose goodwill and favorable opinions are considered essential for a company's free operation. Managerial adeptness at inconsistency becomes evident in the widely discrepant perspectives, reasons for action, and presentations of fact that explain, excuse, or justify managerial behavior to these diverse audiences. (Location 3506)

165. The Supreme Court, however, decided in 1981 that the 1978 OSHA ruling was fully within the Agency's mandate, namely to protect workers' health and safety as the primary benefit exceeding all cost considerations. (Location 3533)

166. The segmented work patterns of bureaucracy underlie these larger structures. Managers' cognitive maps to the thickets of their world contain sharp, sometimes absurd, caricatures of the style and ethos of different occupational groups. These suggest some of the ways in which managers appraise the myriad of character types whom they see peopling their world. Production types, for instance, are said to be hard-drinking, raucous, good-time charlies; engineers, always distinguishable by the plastic pen containers in their shirt pockets, are hostages to an outdated belief in a pristine mathematical rationality; accountants are bean counters who know how to play the shell game; lawyers are legal eagles or legal beagles in wool pinstripes who, if they had their way, would tie managers' hands completely; corporate staff are the king's spies, always ready to do his bidding and his dirty work; marketing guys are cheerful, smooth-talking, upbeat fashion plates who must nonetheless keep salesmen

under their thumbs; salesmen are aggressive loudmouths who feel that they can sell freezers to penguins in Antarctica and who would sell their grandmothers just to make a deal. Salesmen hate the restraints that marketers put on their work and on their ego gratification. Financial wizards, on the ascendancy everywhere, are tight-mouthed, close to the vest poker players who think that a social order can be built on paper deals. And outside consultants are men and women who borrow one's watch and then charge for telling the time. (Location 4327)

167. But so what, they argued, if the preservative did in fact pose some risk of cancer? Better, they said, the risk of a slight long-run increase in the rate of stomach and intestinal cancer than the certainty of a precipitous spurt in the incidence of botulism, particularly in the lower-income black and Hispanic groups that typically consume large amounts of processed meat and, both because of poverty and cultural practices, often leave food uncovered and unrefrigerated for considerable periods. Is corporate social responsibility, they asked, maintaining a private sense and public image of moral purity while someone else does necessary but tainted work? Or is real social responsibility the willingness to get one's hands dirty, to make whatever compromises have to be made to produce a product with some utility, to achieve therefore some social good, even though one knows that one's accomplishments and motives will inevitably be misinterpreted by others for their own ends, usually by those with the least reason to complain? Besides, they pointed out, consumers continue to purchase artificially preserved meats in large quantities. Is not the proper role of business "to give the public what it wants," adopting the market as its polar star, as the only reliable guide in a pluralistic society to "the greatest good for the greatest number," as the final arbiter not of values, which are always arguable, but, more importantly, of tastes, about which there can be no reasonable dispute? (Location 4500)

168. In fact, exercises in substantive rationality—the critical, reflective use of reason—are not only subject to infinite interpretations and counterinterpretations but also invite fantastic constructions of reality, including attributions of conspiracy. Thus a major corporation provides a gift of $10 million to establish new foundations that will materially aid South African blacks and is promptly accused by a black American leader of bolstering apartheid. Weft managers create an elaborate recreational complex for Weft employees in the corporation's southern community and are charged with perpetuating traditional textile company paternalism. Some executives at Images Inc. donate their time to bring together several institutional sectors of a local town in which they live for community betterment and are charged with trying to grab headlines and line up future business. Managers often feel that, however genuine it may be, altruism is a motive that is always denied them by others. To complicate matters still further, the necessary self-promotional work of presenting private goals as public goods, or the self-defensive work within the corporation of presenting public goods as hardheaded business decisions, or managers' knowledge that bureaucracy insulates them from the real consequences of their actual choices, often make their protestations of socially responsible actions suspect even to themselves. (Location 4512)

Interpretations of "probability"

(*Written for Arbital in 2016.*)

What does it *mean* to say that a flipped coin has a 50% probability of landing heads?

Historically, there are two popular types of answers to this question, the "frequentist" and "subjective" (aka "Bayesian") answers, which give rise to radically different approaches to experimental statistics. There is also a third "propensity" viewpoint which is largely discredited (assuming the coin is deterministic). Roughly, the three approaches answer the above question as follows:

- **The propensity interpretation:** Some probabilities are just out there in the world. It's a brute fact about coins that they come up heads half the time. When we flip a coin, it has a fundamental *propensity* of 0.5 for the coin to show heads. When we say the coin has a 50% probability of being heads, we're talking directly about this propensity.
- **The frequentist interpretation:** When we say the coin has a 50% probability of being heads after this flip, we mean that there's a class of events similar to this coin flip, and across that class, coins come up heads about half the time. That is, the *frequency* of the coin coming up heads is 50% inside the event class, which might be "all other times this particular coin has been tossed" or "all times that a similar coin has been tossed", and so on.
- **The subjective interpretation:** Uncertainty is in the mind, not the environment. If I flip a coin and slap it against my wrist, it's already landed either heads or tails. The fact that I don't know whether it landed heads or tails is a fact about me, not a fact about the coin. The claim "I think this coin is heads with probability 50%" is an *expression of my own ignorance,* and 50% probability means that I'd bet at 1 : 1 odds (or better) that the coin came up heads.

For a visualization of the differences between these three viewpoints, see Correspondence visualizations for different interpretations of "probability". For examples of the difference, see Probability interpretations: Examples. See also the Stanford Encyclopedia of Philosophy article on interpretations of probability.

The propensity view is perhaps the most intuitive view, as for many people, it just feels like the coin is intrinsically random. However, this view is difficult to reconcile with the idea that once we've flipped the coin, it has already landed heads or tails. If the event in question is decided deterministically, the propensity view can be seen as an instance of the mind projection fallacy: When we mentally consider the coin flip, it feels 50% likely to be heads, so we find it very easy to imagine a *world* in which the coin is *fundamentally* 50%-heads-ish. But that feeling is actually a fact about *us,* not a fact about the coin; and the coin has no physical 0.5-heads-propensity hidden in there somewhere — it's just a coin.

The other two interpretations are both self-consistent, and give rise to pragmatically different statistical techniques, and there has been much debate as to which is preferable. The subjective interpretation is more generally applicable, as it allows one to assign probabilities (interpreted as betting odds) to one-off events.

Frequentism vs subjectivism

As an example of the difference between frequentism and subjectivism, consider the question: "What is the probability that Hillary Clinton will win the 2016 US presidential election?", as analyzed in the summer of 2016.

A stereotypical (straw) frequentist would say, "The 2016 presidential election only happens once. We can't *observe* a frequency with which Clinton wins presidential elections. So we can't do any statistics or assign any probabilities here."

A stereotypical subjectivist would say: "Well, prediction markets tend to be pretty well-calibrated about this sort of thing, in the sense that when prediction markets assign 20% probability to an event, it happens around 1 time in 5. And the prediction markets are currently betting on Hillary at about 3 : 1 odds. Thus, I'm comfortable saying she has about a 75% chance of winning. If someone offered me 20 : 1 odds *against* Clinton — they get $1 if she loses, I get $20 if she wins — then I'd take the bet. I suppose you could refuse to take that bet on the grounds that you Just Can't Talk About Probabilities of One-off Events, but then you'd be pointlessly passing up a really good bet."

A stereotypical (non-straw) frequentist would reply: "I'd take that bet too, of course. But my taking that bet *is not based on rigorous epistemology,* and we shouldn't allow that sort of thinking in experimental science and other important venues. You can do subjective reasoning about probabilities when making bets, but we should exclude subjective reasoning in our scientific journals, and that's what frequentist statistics is designed for. Your paper should not conclude "and therefore, having observed thus-and-such data about carbon dioxide levels, I'd personally bet at 9 : 1 odds that anthropogenic global warming is real," because you can't build scientific consensus on opinions."

...and then it starts getting complicated. The subjectivist responds "First of all, I agree you shouldn't put posterior odds into papers, and second of all, it's not like your method is truly objective — the choice of "similar events" is arbitrary, abusable, and has given rise to p-hacking and the replication crisis." The frequentists say "well your choice of prior is even more subjective, and I'd like to see you do better in an environment where peer pressure pushes people to abuse statistics and exaggerate their results," and then down the rabbit hole we go.

The subjectivist interpretation of probability is common among artificial intelligence researchers (who often design computer systems that manipulate subjective probability distributions), Wall Street traders (who need to be able to make bets even in relatively unique situations), and common intuition (where people feel like they can say there's a 30% chance of rain tomorrow without worrying about the fact that tomorrow only happens once). Nevertheless, the frequentist interpretation is commonly taught in introductory statistics classes, and is the gold standard for most scientific journals.

A common frequentist stance is that it is virtuous to have a large toolbox of statistical tools at your disposal. Subjectivist tools have their place in that toolbox, but they don't deserve any particular primacy (and they aren't generally accepted when it comes time to publish in a scientific journal).

An aggressive subjectivist stance is that frequentists have invented some interesting tools, and many of them are useful, but that refusing to consider subjective probabilities is toxic. Frequentist statistics were invented in a (failed) attempt to keep subjectivity out of science in a time before humanity really understood the laws of probability theory. Now we have theorems about how to manage subjective probabilities correctly, and how to factor personal beliefs out from the objective evidence provided by the data, and if you ignore these theorems you'll get in trouble. The frequentist interpretation is broken, and that's why science has p-hacking and a replication crisis even as all the wall-street traders and AI scientists use the Bayesian interpretation. This "let's compromise and agree that everyone's viewpoint is valid" thing is all well and good, but how much worse do things need to get before we say "oops" and start acknowledging the subjective probability interpretation across all fields of science?

The most common stance among scientists and researchers is much more agnostic, along the lines of "use whatever statistical techniques work best at the time, and use frequentist techniques when publishing in journals because that's what everyone's been doing for decades upon decades upon decades, and that's what everyone's expecting."

See also Subjective probability and Likelihood functions, p-values, and the replication crisis.

Which interpretation is most useful?

Probably the subjective interpretation, because it subsumes the propensity and frequentist interpretations as special cases, while being more flexible than both.

When the frequentist "similar event" class is clear, the subjectivist can take those frequencies (often called base rates in this context) into account. But unlike the frequentist, she can also combine those base rates with other evidence that she's seen, and assign probabilities to one-off events, and make money in prediction markets and/or stock markets (when she knows something that the market doesn't).

When the laws of physics actually do "contain uncertainty", such as when they say that there are multiple different observations you might make next with differing likelihoods (as the Schrodinger equation often will), a subjectivist can combine her propensity-style uncertainty with her personal uncertainty in order to generate her aggregate subjective probabilities. But unlike a propensity theorist, she's not forced to think that *all* uncertainty is physical uncertainty: She can act like a propensity theorist with respect to Schrodinger-equation-induced uncertainty, while still believing that her uncertainty about a coin that has already been flipped and slapped against her wrist is in her head, rather than in the coin.

This fully general stance is consistent with the belief that frequentist tools are useful for answering frequentist questions: The fact that you can *personally* assign probabilities to one-off events (and, e.g., evaluate how good a certain trade is on a prediction market or a stock market) does not mean that tools labeled "Bayesian" are always better than tools labeled "frequentist". Whatever interpretation of "probability" you use, you're encouraged to use whatever statistical tool works best for you at any given time, regardless of what "camp" the tool comes from. Don't let the fact that you think it's possible to assign probabilities to one-off events prevent you from using useful frequentist tools!

Which scientific discovery was most ahead of its time?

Looking into the history of science, I've been struck by how continuous scientific progress seems. Although there are many examples of great intellectual breakthroughs, most of them build heavily on existing ideas which were floating around immediately beforehand - and quite a few were discovered independently at roughly the same time (see https://en.m.wikipedia.org/wiki/List_of_multiple_discoveries).

So the question is: which scientific advances were most ahead of their time, in the sense that if they hadn't been made by their particular discoverer, they wouldn't have been found for a long time afterwards? (Ideally taking into account the overall rate of scientific progress: speeding things up by a decade in the 20th century seems about as impressive a feat as speeding things up by half a century in ancient Greece).

Where are people thinking and talking about global coordination for AI safety?

Many AI safety researchers these days are not aiming for a full solution to AI safety (e.g., the classic Friendly AI), but just trying to find good enough partial solutions that would buy time for or otherwise help improve global coordination on AI research (which in turn would buy more time for AI safety work), or trying to obtain partial solutions that would only make a difference if the world had a higher level of global coordination than it does today.

My question is, who is thinking directly about how to achieve such coordination (aside from FHI's Center for the Governance of AI, which I'm aware of) and where are they talking about it? I personally have a bunch of questions related to this topic (see below) and I'm not sure what's a good place to ask them. If there's not an existing online forum, it seems a good idea to start thinking about building one (which could perhaps be modeled after the AI Alignment Forum, or follow some other model).

1. What are the implications of the current US-China trade war?
2. Human coordination ability seems within an order of magnitude of what's needed for AI safety. Why the coincidence? (Why isn't it much higher or lower?)
3. When humans made advances in coordination ability in the past, how was that accomplished? What are the best places to apply leverage today?
4. Information technology has massively increased certain kinds of coordination (e.g., email, eBay, Facebook, Uber), but at the international relations level, IT seems to have made very little impact. Why?
5. Certain kinds of AI safety work could seemingly make global coordination harder, by reducing perceived risks or increasing perceived gains from non-cooperation. Is this a realistic concern?
6. What are the best intellectual tools for thinking about this stuff? Just study massive amounts of history and let one's brain's learning algorithms build what models it can?

Authoritarian Empiricism

This is a linkpost for http://benjaminrosshoffman.com/authoritarian-empiricism/

(Excerpts from a conversation with my friend Mack, very slightly edited for clarity and flow, including getting rid of most of the metaconversation.)

Ben: Just spent 2 full days offline for the holiday - feeling good about it, I needed it.

Mack: Good!

Ben: Also figured out some stuff about acculturation I got and had to unlearn, that was helpful

Mack: I'm interested if you feel like elaborating

Ben: OK, so, here's the deal.

I noticed over the first couple days of Passover that the men in the pseudo-community I grew up in seem to think there's a personal moral obligation to honor contracts, pretty much regardless of the coercion involved. The women seem to get that this increases the amount of violence in the world by quite a lot relative to optimal play, but they don't really tell the men. This seems related somehow to a thing where the men feel anxious about the prospect of modeling people as autonomous subjects - political creatures - instead of just objectifying them, but when they slap down attempts to do that, they pretend they're insisting on rigor and empiricism.

Which I'd wrongly internalized, as a kid, as good-faith critiques of my epistemics.

Story 1:

I was talking with my father about Adorno, the Enlightenment, and anti-Semitism, and the conversation was doing a reasonable-seeming thing, UNTIL he brought up the issue of high-fertility ethnic minorities with distinct political loyalties in democracies. So, naturally, first I explored the specific thing he brought up, which was that this strategy exploits a real security flaw in the democratic setup, and (since this came up in the context of Israel) that hypocritical ethnic majorities willing to occasionally violate their "standards" do a lot better patching the security flaw, than do ethnic majorities who insist on ACTUALLY having structurally neutral liberalism that takes care of and empowers everyone.

But, then, since we'd been talking about anti-Semitism, I had to point out that there's a structurally similar thing going on with Jews and credit-allocation systems in early financialized states like pre- and interwar Germany. If there had been actual coordination and an actual agenda, it would have been trivial to take over the state. (There wasn't and there wasn't, it's a trope in pre-WWII-era Jewish humor that the anti-Semitic newspapers kind of read like escapist fantasy). But, like, a double-digit percentage of elites is obviously enough, in a modern state where info-processing is abstract and mostly automated, to control quite a lot, given perfect coordination.

And he basically said, "you can't say that, because you don't have hard data."

Which, like, where am I gonna find hard data on the incidence of coups via groups with unreasonably high levels of coordination seizing control of the state's information-processing apparatus (thus causing the records to misreport reality as a side effect)?

When I poked him on this, he ended up retreating to the motte of "it's possible that what you're saying isn't true". Which, yes, obviously - it's speculation. But also obviously that isn't what he was originally saying. He was saying something like: It's wrong to reason about concrete situations based on hypotheticals about human potential; legitimate discourse is the sort of thing that could get into an academic journal (which is necessarily at least performing being apolitical in some sense, even in the journal's explicitly about political theory).

This helped a bunch of past stuff click for me, where e.g. he knows a lot about what the Rabbis of the Talmud said, and what later medieval commentators have said, and historical scholarship about how the text developed, and that's fine to talk about, but if I read them as though they were arguing about some specific real thing, try to understand and then talk about it, and use it to contextualize individual statements, that seems like "irresponsible" speculation to him.

Digression to an example I think is cool:

At the Passover Seder, we traditionally read a story about five Rabbis, in Roman times, staying up all night to study the Exodus from Egypt (the Passover story). (These are guys who were also associated both with the rebellions against Rome, and the successful transition to a permanently exilic Judaism.) And then in the morning their students come in and say "it's time to recite the morning Shema" (central affirmation that traditional Jews recite communally twice daily).

Turns out there's ANOTHER story in the Talmud about a rabbi staying up studying until his students come in to tell him it's time for the morning Shema, but this one is very different. It's Bar Yochai, a figure associated with mysticism / proto-Kabbalah. He's just generically studying Torah, not specifically the Exodus story. He's alone, not with peers. And when his students come in, he says that studying Torah takes precedence over anything else, so he's not going to come say the Shema with them, even though it's an obligatory commandment.

This is part of a broader disagreement between Bar Yochai and the other rabbis.

Another instance of the same disagreement:
Most of the Rabbis think that the commandment to attend to Torah (the teachings of Moses) all day means that if e.g. you're planting your crops, figure out how to do that in a Torah-ish way. Bar Yochai says you should literally just sit studying Torah, and if you do that well enough, gentiles will show up and plant your crops for you as a reward. So, Bar Yochai and his students tried it his way, and the other rabbis and their students tried it their way. And, empirically, Bar Yochai turned out to be mistaken. He got magic powers (the Talmud is very clear on this point), but his crops failed because he ... didn't plant or harvest them.

Basically he prioritized inner work over everything else, assuming that it's high enough leverage that other stuff would take care of itself, and the other rabbis thought that this stuff doesn't work outside the context of a community operating with some sort of synchronization, or outside the context of the mundane activities of life.

It's not hard to see why (a) the Talmud says that if there's any other school of thought available, never go with Bar Yochai's opinion on a legal matter, but also (b) the kabbalists saw him as an intellectual precursor.

So, linking this back to the underlying problem - describing the stories is OK, making inferences about them sort of registers as a kind of storytelling that can be fun/interesting, but my dad just can't engage with the idea that there's a fact of the matter about what these people were talking *about*, separate from what they explicitly said, and talk about kabbalah as political theory of change with concrete mundane implications.

Story 2:
I'd just talked with my mom a bunch about her adult ESL students - some of them are "unmotivated" and she'd recently realized it's in part because some are coerced to show up lest they lose their visas. I pointed out that she could just negotiate directly with them to work out a solution that allows the ones who want to learn to not be distracted, and that she's not morally obliged to force the ones who aren't interested in the class to pretend they are.

Then at 2nd seder a friend's father was talking with her about this, and as soon as he heard about the symptoms, he declared that she should set a firm boundary so that students that e.g. after n minutes the door of the classroom is locked and students who are too late are absent, that her first obligation is to her contract as a teacher, etc. And he basically just couldn't hear or wasn't interested in the fact that some of the students were under coercion, didn't seem to think that fact was morally relevant at all.

(None of these examples is hugely persuasive on their own, but each of them caused a long pattern of similar things to click).

When I pointed out that my mom wasn't morally obliged to collaborate with ICE he just denied that this had anything to do with what he was saying, without offering an argument.

Story 2b:
Same night, different incident.

My friend (the son of the guy from story 2) asked me how Pittsburgh was.

I responded with the following analogy:

While in Berkeley, it's like I was living on the first-class deck of the Titanic. In the distance, I can see the ship heading towards an iceberg. Meanwhile, all the first-class passengers are obsessed with scheming about how to become the captain, or otherwise take over the ship and get the nice staterooms and privileges.

I'm concerned with steering the ship to safety, but when I find people rallying around the stated intent to steer the ship to safety, they're mostly just another faction trying to take over the ship. I try to persuade individuals that ACTUALLY navigating is object-level important even though it doesn't affect anything in our immediate concrete environment, but this just seems to people like a weird bank-shot attempt to gain status by dominating the "steer the ship to safety" faction.

So, depressed and scared and emotionally scarred by this, I go to a place I've heard there are a bunch of sane competent engineers: the engine room!

It turns out, they ARE locally sane here. They're collaborating to do means-ends reasoning to keep the engine running, which keeps the lights on and keeps the ship moving forwards. Given the crazy situation we're in, keeping the ship moving forwards is not helping. But at least it's literally not their job to know about that, and they're doing what literally is their job. When I describe what's going on on the upper deck they don't seem particularly inclined to drop everything and come help, but they do seem sincerely concerned and interested in finding out whether they have any relevant resources they can direct to me. They understand in principle why steering the ship matters, and that hitting an iceberg would be bad in a way totally unrelated to factional politics.

Pittsburgh is the engine room.

So, I'm in the part of this analogy that's about the Bay, and my friend's dad jumps into the conversation to tell me that my analogy is too convoluted. So, I pause and ask him what part's hard to follow (he wasn't part of the conversation at first, but if someone wants to understand what I'm saying at a social event, it seems correct to try to include them), and he just keeps repeating that it's too convoluted, until eventually he changes his story and says "it's too crazy, I don't want to hear about it."

So, he was pretending to be critiquing my analogy, actually feels too much anxiety about the situation I'm describing to be OK letting someone else talk about it where he can hear, but felt the need to put himself above me by framing it as me making some sort of technical error in conversation.

Do you see how this seems like the same kind of thing my actual dad did?

Meanwhile, (back to the contracts thing), his wife works as a lawyer to advocate for kids whose needs aren't met by the family law & school system. She can't possibly do that job and think that the letter of the law even has an objective meaning, since it's literally her job to make it mean the thing that gets an okay outcome for the child.

The men of this category often end up in a position where they are the only one in their area who are technically adept at the thing people with their job description are supposedly certified to know about, or who care to do the object-level technical work.

I think this specific gendered dynamic might be particular to secular American Jews.

Mack: Okay that's interesting. Definitely seen similar things play out but not in such a gendered way. Thinking about my parents in particular, they end up on the "male" side of your stories occasionally. Not consistently at all. Hm maybe the examples coming to mind are only superficially similar.

Ben: Want to work through the details of one? Might be good to precisely formulate the distinction if there is one.

Mack: Re: not treating people like political entities, I can think of examples of that. But I suspect the reasons are different.

Ben: I suspect there's a shared sense to think of people of the other political party as defective parts of a machine, rather than as adversaries who might be negotiated with or fought but with whom there's not currently a shared paradigm. But, not a shared tendency to specifically dismiss attempts to model people as agents, as unscientific.

Mack: Things that come to mind: a knee jerk reaction among the older members on one side of the family to treat this kind of reasoning as...vulgar?

Ben: What does an example of the sort of thing they've reacted to this way look like? Actual or fictional examples both fine. Actual are better, but whatever prediction/generation function you learned is also valuable intel. (Just like fictional stories by competent poets are valuable intel)

Mack: I'm thinking of a cousin who is very similar to me. There are running jokes about us being in the same room and driving people crazy because we "start controversies." I think there was a conversation about Boise's homelessness policies, and she and I were talking about things like: the reasons the city might have taken recent aggressive action against the homeless population, essentially the different incentives at play.

We disagreed but it was sane disagreement, and her mother and grandmother were just visibly distressed. And they tried talking about ministry attempts, harsher drug laws, etc. Retreating to party lines on homelessness (red tribe). The conversation ended with her mother saying "Well then why bother!" as we poked at the policies they'd brought up.

It isn't the same retreat to what could be published in an academic journal, or to the obligation of a contract. But it is kind of like your Titanic analogy. Laying out the specific reasons a problem is hard, the normal party lines or grumbling not being sufficient or satisfying, and finding it rude to point out why a problem is hard, especially if it isn't about the outgroup being wrong or misled by satanic forces.

Ben: OK, so it sounds like your family is nondissociatedly anxious about politics, while mine (at least the men) retreats to dissociatedly identifying with an authority narrative that insists that only "apolitical" knowledge is speakable; your family more overtly identifies as members of a faction, while mine identifies with abstract shared authority.

Mack: That sounds right. Ah, so this side of the family is also pretty bound to contracts of a sort, though they aren't quite as aligned with the law.

Ben: All "legally binding" contracts, or just uncoerced personal agreements?

Mack: All legally binding contracts to an extent though that's more about avoiding punishment and being Good. Seeing the local social mores *as* binding contracts, I think.

Ben: That last thing seems noncrazy to me - like, an attitude I'd see in some fully functional societies.

Mack: It isn't crazy.

Ben: Whereas I think the thing I was pointing to is crazy, and the other things are somewhere in between.

Mack: I think I see the distinction. Feels like there's something familiar in my experience that's closer to the crazy side and I'm trying to figure out where that comes from.

Initial recognition was about the discomfort and retreat - I have a lot of examples of the role you took in those anecdotes being seen as extremely rude, uncomfortable, vulgar. I don't think it comes from the same place as the specific dynamic, though.

Recognition also of the realization that the people arguing around me were not arguing to try to understand something or solve the problem.

Ben: OK, I think the discomfort-and-retreat pattern is a specific kind of defensiveness, on behalf of the ruling regime by people identifying with it (where the ruling regime can be a local community's norms, or the state, or an ideology, etc etc.)

That's an important piece of model to have, it's one of the gears here. It connects to more than one possible type of defense or sense of threat.

Mack: I am curious about whether I've observed something closer.

Maybe this: at work some very expensive material was mixed. The timeline for new material was too long to meet even the revised deadlines for the product, there was no good mechanical solution, etc. So the bosses had been rotating employees through the tedious task of unmixing it by hand.

HR lady and I helped with this during some plant wide mandatory overtime.

She was insistent that the right thing to do would be to force the person responsible for the mess to devote all of their work hours plus overtime to fixing it.

I argued a little - not too hard because office norms. But her retreat was to a supposed alignment with company interests (even though, IMO, the solution was an okay compromise with multiple goals for the plant).

And it has come out over time that, as far as I can tell, she believes very strongly that when you begin employment you must suspend a large chunk of your personal interests and align them with the firm, or you're a subpar employee. And while this probably helps her a lot in some of her HR functions, she is resistant to discussing the individual incentives that prevent people from being "good employees" once they've come on to her radar as "bad employees."

Ben: The HR thing sounds like it might be an exact match with a big part of this. I do want to distinguish loyalty to a specific local institution, from loyalty to one's profession/contract. They're different kinds of implied coordination strategies.

Mack: Which loyalty is the one present in your stories?

Ben: The latter. So, the HR lady identifies her interests with the interests of the company she's attached to, that's her gang. But the guys I'm talking about identify with each other as members of a mercenary class with a perceived shared interest in upholding professional standards, so that they can be interchangeable pieces and charge for this.

Mack: Ahhh
Okay
Something clicked

Ben: Like, a doctor will identify with Doctors as a profession, not with the hospital and nurses. In-house counsel will often favor the class interests of lawyers over the

interests of their company.

The Guild.

*****

Related: Blind Empiricism

Probability interpretations: Examples

(*Written for Arbital in 2016.*)

Betting on one-time events

Consider evaluating, in June of 2016, the question: "What is the probability of Hillary Clinton winning the 2016 US presidential election?"

On the **propensity** view, Hillary has some fundamental chance of winning the election. To ask about the probability is to ask about this objective chance. If we see a prediction market in which prices move after each new poll — so that it says 60% one day, and 80% a week later — then clearly the prediction market isn't giving us very strong information about this objective chance, since it doesn't seem very likely that Clinton's *real* chance of winning is swinging so rapidly.

On the **frequentist** view, we cannot formally or rigorously say anything about the 2016 presidential election, because it only happens once. We can't *observe* a frequency with which Clinton wins presidential elections. A frequentist might concede that they would cheerfully buy for $1 a ticket that pays $20 if Clinton wins, considering this a favorable bet in an *informal* sense, while insisting that this sort of reasoning isn't sufficiently rigorous, and therefore isn't suitable for being included in science journals.

On the **subjective** view, saying that Hillary has an 80% chance of winning the election summarizes our *knowledge about* the election or our *state of uncertainty* given what we currently know. It makes sense for the prediction market prices to change in response to new polls, because our current state of knowledge is changing.

A coin with an unknown bias

Suppose we have a coin, weighted so that it lands heads somewhere between 0% and 100% of the time, but we don't know the coin's actual bias.

The coin is then flipped three times where we can see it. It comes up heads twice, and tails once: HHT.

The coin is then flipped again, where nobody can see it yet. An honest and trustworthy experimenter lets you spin a wheel-of-gambling-odds — reducing the worry that the experimenter might know more about the coin than you, and be offering you a deliberately rigged bet — and the wheel lands on (2 : 1). The experimenter asks if you'd enter into a gamble where you win $2 if the unseen coin flip is tails, and pay $1 if the unseen coin flip is heads.

On a **propensity** view, the coin has some objective probability between 0 and 1 of being heads, but we just don't know what this probability is. Seeing HHT tells us that

the coin isn't all-heads or all-tails, but we're still just guessing — we don't really know the answer, and can't say whether the bet is a fair bet.

On a **frequentist** view, the coin would (if flipped repeatedly) produce some long-run frequency f of heads that is between 0 and 1. If we kept flipping the coin long enough, the actual proportion p of observed heads is guaranteed to approach f arbitrarily closely, eventually. We can't say that the *next* coin flip is guaranteed to be H or T, but we can make an objectively true statement that p will approach f to within epsilon if we continue to flip the coin long enough.

To decide whether or not to take the bet, a frequentist might try to apply an unbiased estimator to the data we have so far. An "unbiased estimator" is a rule for taking an observation and producing an estimate e of f, such that the expected value of e is f. In other words, a frequentist wants a rule such that, if the hidden bias of the coin was in fact to yield 75% heads, and we repeat many times the operation of flipping the coin a few times and then asking a new frequentist to estimate the coin's bias using this rule, the *average* value of the estimated bias will be 0.75. This is a property of the *estimation rule* which is objective. We can't hope for a rule that will always, in any particular case, yield the true f from just a few coin flips; but we can have a rule which will provably have an *average* estimate of f, if the experiment is repeated many times.

In this case, a simple unbiased estimator is to guess that the coin's bias f is equal to the observed proportion of heads, or 2/3. In other words, if we repeat this experiment many many times, and whenever we see p heads in 3 tosses we guess that the coin's bias is $\frac{p}{3}$, then this rule definitely is an unbiased estimator. This estimator says that a bet of $2 vs. $1 is fair, meaning that it doesn't yield an expected profit, so we have no reason to take the bet.

On a **subjectivist** view, we start out personally unsure of where the bias f lies within the interval [0, 1]. Unless we have any knowledge or suspicion leading us to think otherwise, the coin is just as likely to have a bias between 33% and 34%, as to have a bias between 66% and 67%; there's no reason to think it's more likely to be in one range or the other.

Each coin flip we see is then evidence about the value of f, since a flip H happens with different probabilities depending on the different values of f, and we update our beliefs about f using Bayes' rule. For example, H is twice as likely if $f = \frac{2}{3}$ than if $f = \frac{1}{3}$ so by Bayes's Rule we should now think f is twice as likely to lie near $\frac{2}{3}$ as it is to lie near $\frac{1}{3}$.

When we start with a uniform prior, observe multiple flips of a coin with an unknown bias, see M heads and N tails, and then try to estimate the odds of the next flip coming up heads, the result is Laplace's Rule of Succession which estimates (M + 1) : (N + 1) for a probability of ~~M+1N+2~~.

In this case, after observing HHT, we estimate odds of 2 : 3 for tails vs. heads on the next flip. This makes a gamble that wins $2 on tails and loses $1 on heads a profitable gamble in expectation, so we take the bet.

Our choice of a uniform prior over f was a little dubious — it's the obvious way to express total ignorance about the bias of the coin, but obviousness isn't everything. (For example, maybe we actually believe that a fair coin is more likely than a coin biased 50.0000023% towards heads.) However, all the reasoning after the choice of prior was rigorous according to the laws of probability theory, which is the only method of manipulating quantified uncertainty that obeys obvious-seeming rules about how subjective uncertainty should behave.

Probability that the 98,765th decimal digit of π is 0

What is the probability that the 98,765th digit in the decimal expansion of π is 0?

The **propensity** and **frequentist** views regard as nonsense the notion that we could talk about the *probability* of a mathematical fact. Either the 98,765th decimal digit of π is 0 or it's not. If we're running *repeated* experiments with a random number generator, and looking at different digits of π, then it might make sense to say that the random number generator has a 10% probability of picking numbers whose corresponding decimal digit of π is 0. But if we're just picking a non-random number like 98,765, there's no sense in which we could say that the 98,765th digit of π has a 10% propensity to be 0, or that this digit is 0 with 10% frequency in the long run.

The **subjectivist** considers probabilities to just refer to their own uncertainty. So if a subjectivist has picked the number 98,765 without yet knowing the corresponding digit of π, and hasn't made any observation that is known to them to be entangled with the 98,765th digit of π, and they're pretty sure their friend hasn't yet looked up the 98,765th digit of π either, and their friend offers a whimsical gamble that costs $1 if the digit is non-zero and pays $20 if the digit is zero, the Bayesian takes the bet.

Note that this demonstrates a difference between the subjectivist interpretation of "probability" and Bayesian probability theory. A perfect Bayesian reasoner that knows the rules of logic and the definition of π must, by the axioms of probability theory, assign probability either 0 or 1 to the claim "the 98,765th digit of π is a 0" (depending on whether or not it is). This is one of the reasons why perfect Bayesian reasoning is

intractable. A subjectivist that is not a perfect Bayesian nevertheless claims that they are personally uncertain about the value of the 98,765th digit of π. Formalizing the rules of subjective probabilities about mathematical facts (in the way that probability theory formalized the rules for manipulating subjective probabilities about empirical facts, such as which way a coin came up) is an open problem; this in known as the problem of logical uncertainty.

Towards optimal play as Villager in a mixed game

This is a linkpost for http://benjaminrosshoffman.com/towards-optimal-play-as-villager-in-a-mixed-game/

On Twitter, Freyja wrote:

> *Things capitalism is trash at:*
> *Valuing preferences of anything other than adults who earn money (i.e. future people, non-humans)*
> *Pricing non-standardisable goods (i.e. information)*
> *Playing nicely with non-quantifiable values + objectives (i.e. love, ritual)*
>
> *Things capitalism is good at:*
> *Incentivising the production of novel goods and services*
> *Coordinating large groups of people to produce complex bundles of goods*
> *The obvious: making value fungible*
>
> *Anyone know of work on -*
> *a) integrating the former into existing economic systems, or*
> *b) developing new systems to provide those things while including capitalism's existing benefits?*

This intersected well enough with my current interests and those of the people I've been discoursing with most closely that I figured I'd try my hand at a quick explanation of what we're doing, which I've lightly edited into blog post form below. This is only a loose sketch, I think it does reasonably precisely outline the argument, but many readers may find that there are substantial inferential leaps. Questions in the comments are strongly encouraged.

Any serious attempt at (b) will first have to unwind the disinformation that claims that the thing we have now is capitalism, or remotely efficient.

The short version of the project: learning to talk honestly within a small group about how power works, both systemically and as it applies to us, without trying to hold onto information asymmetries. (There's pervasive temptation to withhold political information as part of a zero-sum privilege game, like Plato's philosopher-kings.)

Some background: post-WWII elite institutions (e.g. corps) are competitive to enter, but not under performance pressure, because of US government policy. This strongly selects for zero-sum games, which mimic but wreck discourse. (See Moral Mazes for more, especially the case studies that make up most of the book, starting around chapter 3.)

This creates opportunity in two ways.

First, institutions are mostly too stupid to model their environment beyond the zero-sum games they specialize in, so a small group that's able to maintain information hygiene and not turn on each other should be able to take & hold territory. "And not turn on each other" turns out to be really hard, because all our role models and

intuitions for how to survive in this world involve doing that all the time. But we're learning!

(A mundane example of a decisive advantage due to information hygiene: Paul Graham writes about how his startup did better because it used an elegant programming language. That's only information hygiene on the purely technical level, but that was enough to outmaneuver huge corporations with a strong perceived incentive to ruin them, for quite a while. For a less mundane example, the story of how Elisha outmaneuvered multiple ruling dynasties is a personal favorite - 2 Kings 5-10. The narrative distorts the "miracles" a bit but it's not hard to reconstruct how he actually did it.)

Second, because most supposed productive activity is done in the context of huge stable corporations, people are trying to maximize the number of jobs and complexity per unit of output. This implies that many things can be done much more easily.

So that implies that if we can have good enough information hygiene and group cohesion not to fall victim to the perverse impulse to do the kind of make-work or artificial scarcity that creates much of cost disease, we can learn how to build a nearly full-stack civilization in a small city-state. Obviously there are many steps between here and there, but since lots of them involve getting collectively smarter, a detailed plan would be inappropriate.

What does good information hygiene and group cohesion look like? The game Werewolf is a good example. Players are secretly assigned the identity of Villager (initially the majority) or Werewolf (minority). Each round all players vote one player out, and Werewolves secretly do the same. There are other details that allow villagers to make some inferences about who the werewolves are. But they have to play the first few rounds right or they lose.

Optimal play for Werewolves involves (a) targeting whichever villagers are the most helpful to public deliberation, for exclusion, and (b) during public deliberation, being as unhelpful as they can get away with while appearing to try to help at other times. I realized a lot of things about how social skills feel from the inside when I finally figured out how to play correctly as a Werewolf.

Optimal play for Villagers involves creating as much clarity as possible, as soon as possible, and being willing to assume that people who seem to be foolishly gumming up the works are Werewolves if there's no other clear target.

With optimal play, Villagers usually win, but in practice, at best one or two people try to create clarity and are picked off in the first round by the Werewolves. The other Villagers are resigned to trying to die last, so they lose.

The thing I said about elite culture favoring zero-sum games can be recast as: the social environment favors playing Werewolf over playing Villager. In case it's not obvious, optimal real-world play for Villagers can often involve leaving the Werewolves alone. In real life there are better things to do than murder your enemies, like hang out. Villagers just need to defend themselves if and when they're actually threatened.

We're trying to learn how to play the Villager strategy successfully, in a context where we've mostly been acculturated to play as Werewolves, especially among elites. This has to involve figuring out how to do interpersonal fault analysis (identify when people are being Werewolfy) without scapegoating (assuming that fault -> blame -> exclusion).

In other words, justice seeks truth, but intends to leave no one behind; people who can't contribute need to feel safe admitting that, and people who hurt the group need the option to repent & heal the breach.

We don't have great finesse yet but optimal play in our world seems to be some fluid integration of talking about politics, healing personal trauma, and intersubjective openness.

Havel's The Power of the Powerless describes a similar (but less self-aware) strategy which he calls "dissidence." He (accurately, I think) predicts that the situation in Capitalist countries will be more difficult than the situation in Communist ones, because Capitalist ideology is more persuasive because it's more plausibly true.

Ed Boyden on the State of Science

I just listened to Tyler Cowen's interview with Ed Boyden (link and transcript). The second half contained a lot of questions about current scientific infrastructure, and Boyden had a lot of interesting comments, so I've reproduced a few particular quotes here and added headings.

(Things I've not quoted that LWers might be interested in: Boyden said that whole brain emulations probably work in principle, that he meditates every day and has used an internal family systems meditation for 10 years.)

The Surprisingly Poor State of Funding

COWEN: How should we improve the funding of science in this country?

BOYDEN: I like to look at the history of science to learn about its future, and one thing I've learned a lot over the last couple years—and it's even happened to me—is that it's really hard to fund pioneering ideas.

Brian Kobilka, who recently won the Nobel Prize for solving the structure of the G-protein-coupled receptor—and for context, one-third of all drugs target this class of molecules, so it's a very, very important class of drugs—he lost his funding because he wasn't making progress fast enough. If I recall, he had to moonlight as an emergency room physician to keep going on his research.

Doug Prasher, who cloned the gene for green fluorescent protein, which has been used in something like a million biology studies, ballpark—he lost his funding and eventually left science, ended up driving a shuttle bus for, I believe, a rental car facility or something.

Anyway, there's so many stories. For me, it became personal because when we proposed this expansion microscopy technology, where we blow up brain specimens and other specimens a hundred times in volume to map them, people thought it was nonsense. People were skeptical. People hated it. Nine out of my first ten grants that I wrote on it were rejected.

If it weren't for the Open Philanthropy Project that heard about our struggles to get this project funded—through, again, a set of links that were, as far as I can tell, largely luck driven—maybe our group would have been out of business. But they came through and gave us a major gift, and that kept us going.

> *Boyden also said this sentence in passing, which seemed striking to me about the insularity at the highest echelons of science.*

BOYDEN: I read a statistic that 40 percent of the professors at MIT trained at one point in their career at Stanford, Harvard, or MIT.

How to Improve Funding

COWEN: Let's say you had $10 billion or $20 billion a year, and you would control your own agency, and you were starting all over again, but current institutions stay in place. What would you do with it? How would you structure your grants? You're in charge. You're the board. You do it.

BOYDEN: Yeah, three thoughts. The first thing that I thought a lot about—studying these past cases and then going through it myself—is thinking about peer review. What is peer review?

When you propose a project, a bunch of your peers will then critique it. The problem that a lot of these daring-sounding projects encounter is that they sound bad during peer review because they're so off the wall, or they bring together multiple fields that maybe nobody's qualified to evaluate them.

One thought is, what if—instead of taking people's opinions and then just sort of combining those opinions, and then, okay, you're in or you're out in terms of getting the money—what if we take a step back, and we think about why the peers are thinking this way?

If somebody critiques a proposal, but they're doing it from a vantage point that doesn't see a certain part of the proposal as valuable because they're missing an underlying piece of knowledge, or they're evaluating a proposal—based upon opinion—that, if we think about the logical underpinnings of it, the rationale is actually pretty solid in terms of its being linked to ground-truthable sciences, like physics and chemistry.

In other words, if we take a step back and apply more logical principles of evaluation to the outcomes of peer review, can we actually improve the ranking of these proposals? This is something I'm thinking a lot about right now. As I evaluate people and evaluate ideas that people propose to me as well, I'm trying to hone those skills in myself. That's one of the three things I would do.

[…]The second thing I would do is to be more dynamic in my funding. Right now, maybe there's a grant that you apply for, and then a year later you get the money.

But what if somebody tries something out one Friday afternoon, and whoa, that could cure disease, or that could yield an amazing new insight into biology, or that could allow us to diagnose brain diseases early, or whatever? Why wait a year? What if one could dynamically allocate funding up and down based upon the real-time metrics of science?

In my own group, sometimes we get a project out of the blue, and hey, that's pretty cool. Then we'll dynamically try to understand if we can reallocate resources. That's another thing I would do.

The third thing I would do is I would go looking for trouble. I would go looking for serendipity. If you look at CRISPR for genome editing—that was found by some scientists working on yogurt. If you look at fluorescent proteins—that was identified by a person who just was obsessed with jellyfish.

In my own field, if you look at our optogenetics work or our expansion microscopy work—these fields owe a debt to basic curiosity about critters living in bodies of water for optogenetics, and expansion microscopy goes back to the 1980s where people were wondering why do certain polymers swell so hugely, with no practical-purpose implications of it.

One idea is, how do we find the diamonds in the rough, the big ideas but they're kind of hidden in plain sight? I think we see this a lot. Machine learning, deep learning, is one of the hot topics of our time, but a lot of the math was worked out decades ago—backpropagation, for example, in the 1980s and 1990s. What has changed since then is, no doubt, some improvements in the mathematics, but largely, I think we'd all agree, better compute power and a lot more data.

So how could we find the treasure that's hiding in plain sight? One of the ideas is to have sort of a SWAT team of people who go around looking for how to connect the dots all day long in these serendipitous ways.

COWEN: Does that mean fewer committees and more individuals?

BOYDEN: Or maybe individuals that can dynamically bring together committees. "Hey, you're a yogurt scientist that's curious about this weird CRISPR molecule you just found. Here's some bioinformaticists who are looking to find patterns. Here's some protein engineers who love—"

COWEN: But should the evaluators be fewer committees and more individuals? The people doing the work will always be groups, but committees, arguably, are more conservative. Should we have people with more dukedoms and fiefdoms? They just hand out money based on what they think?

BOYDEN: A committee of people who have multiple non-overlapping domains of knowledge can be quite productive.

What if I brought together to evaluate a proposal, and I have a physicist who can tell me, "You know what? That amount of energy won't kill the brain." Then I have a biologist who says, "You know what? That's a really important problem." And then a chemist who would say, "You know what? That molecule probably won't be toxic." You actually need a committee to judge some of these ideas

Why is Science Slowing Down?

COWEN: Is progress in science slowing down right now?

BOYDEN: That's a good question. I think what's happening is we're tackling bigger problems. Let me explain what that means.

In physics, there's a small number of building blocks, like protons and electrons, and a small number of ways they interact, like electromagnetism and so forth. Chemistry—there's more stuff. There's a hundred-odd things in the periodic table, although maybe there's only 30 to 50 that you actually have to work with if you're trying to make something actually happen. Again, there's a small number of bonds: covalent and ionic and so forth.

I think the problem right now is that a lot of the scientific questions we're wrestling with, whether it's in biology and medicine—but I'm not an expert in this; you know more about some of these things than I do—but in economics and education and so forth, it also seems like—from my distant view—some of these problems relate to this idea that there's a lot of different building blocks and a lot of ways they interact.

In biology, we have what, 30,000 genes in the human genome, and while we know their sequence, for the most part, we have no idea how these gene products interact

with each other, and how they're architected into cells and tissues and organs, and how those go wrong. The problem is this cognitrone explosion of possibilities is so staggeringly huge that a lot of what we try will fail.

What do we do about it? One point of view is, "Well, if we had better tools, and we could map those building blocks and those interactions, maybe we could reduce the risk of biomedical science." Again, it's not my field. You know more about this than I do. I'd love to hear your opinion. But in economics and in other fields, it also seems like people are trying to make better maps of things and how they interact.

That's one idea. What if we could make these problems . . . Progress might seem to be slower because the problems are so hard. But with better tools, maybe we can level the playing field and make 21st-century sciences more tractable, in the same way that 20th-century sciences gave us lasers and computers and the internet.

COWEN: In economics, we have more good empirical papers than ever before, but virtually no more theoretical breakthroughs, and I'm not sure we'll ever have them again.

BOYDEN: Oh, how interesting.

COWEN: That may just be diminishing returns. There are so many fundamental ideas, and you learn those, and you stop, and then you measure things.

BOYDEN: Hmm. Well, in biomedicine, systems didn't evolve to be understood. They evolved to survive and reproduce and all that. One can hope for structure. Biology does give you more structure than we deserve, I think. DNA has a double helix, and you can read out the genetic code.

There's always this question of why is the universe understandable in the first place, and maybe now we're entering the realm of complexity where things are less understandable. But again, we have to accept reality for what it is.

How to Hire Good Scientists

BOYDEN: ...in our group at MIT, I have two PhD students. Neither finished college, actually. I can't think of any other neuroscience groups on Earth where that's true.

Later in the interview.

COWEN: What kind of students are you likely to hire that your peers would not hire?

BOYDEN: Well, I really try to get to know people at a deep level over a long period of time, and then to see how their unique background and interests might change the field for the better.

I have people in my group who are professional neurosurgeons, and then, as I mentioned, I have college dropouts, and I have people who . . . We recently published a paper where we ran the brain expansion process in reverse. So take the baby diaper polymer, add water to expand it, and then you can basically laser-print stuff inside of it, and then collapse it down, and you get a piece of nanotechnology.

The co–first author of that paper doesn't have a scientific laboratory background. He was a professional photographer before he joined my group. But we started talking,

and it turns out, if you're a professional photographer, you know a lot of very practical chemistry. It turns out that our big demo—and why the paper got so much attention—was we made metal nanowires, and the way we did it was using a chemistry not unlike what you do in photography, which is a silver chemistry.

I really try to understand how individual people and their unique background and interests could change the world, but it means that we don't really have a formula. I try not to have formulas, in general, when it comes to the actual day-to-day of science. I often say to people in my group, "We want to revolutionize the world for the better and do the right thing and be ethical, but beyond that, let's not try to make any artificial policies."

How to Find Good Ideas

COWEN: [H]ow do you use discoveries from the past more than other scientists do?

BOYDEN: One way to think of it is that, if a scientific topic is really popular and everybody's doing it, then I don't need to be part of that. What's the benefit of being the 100,000th person working on something?

So I read a lot of *old* papers. I read a lot of things that might be forgotten because I think that there's a lot of treasure hiding in plain sight. As we discussed earlier, optogenetics and expansion microscopy both begin from papers from other fields, some of which are quite old and which mostly had been ignored by other people.

I sometimes practice what I call failure rebooting. We tried something, or somebody else tried something, and it didn't work. But you know what? Something happened that made the world different. Maybe somebody found a new gene. Maybe computers are faster. Maybe some other discovery from left field has changed how we think about things. And you know what? That old failed idea might be ready for prime time.

With optogenetics, people were trying to control brain cells with light going back to 1971. I was actually reading some earlier papers. There were people playing around with controlling brain cells with light going back to the 1940s. What is different? Well, this class of molecules that we put into neurons hadn't been discovered yet.

COWEN: The same is true in economics, I think. Most of behavioral economics you find in Adam Smith and Pigou, who are centuries old.

BOYDEN: Wow. I almost think search engines like Google often are trying to look at the most popular things, and to advance science, what we almost need is a search engine for the most important unpopular things.

COWEN: Sometimes I try doing searches. I take the words I want, and then I throw in a random word that is not related at all, and I try googling that, or through Google Scholar, and I see what comes up.

BOYDEN: Absolutely. I do that a lot, too. That's one thing where I really value those six years I spent learning a bit of chemistry and a bit of physics and a bit of electrical engineering, because it allows me to stitch together some facts from different fields, and that can be very helpful for launching a new idea or judging whether an idea's actually worth pursuing.

In Summary

COWEN: Last question. As a researcher, what could and would you do with more money?

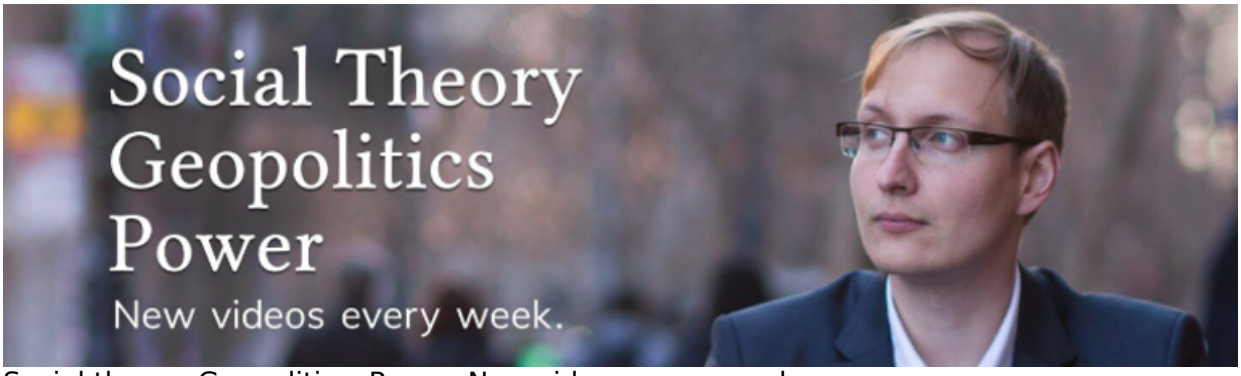
BOYDEN: Well, I'm always looking for new serendipitous things, connecting the dots between different fields. These ideas always seem a bit crazy and are hard to get funded. I see that both in my group but also in many other groups.

I think if I was given a pile of money right now, what I would like to do is to find a way—not just in our group but across many groups—to try to find those unfundable projects where, number one, if we think about the logic of it, "Hey, there's a non-zero chance it could be revolutionary." Number two, we can really, in a finite amount of time, test the idea. And if it works, we can dynamically allocate more money to it. But if it doesn't work, then we can de-allocate money to it.

If I think about optogenetics or expansion microscopy, or these other techniques that we've been talking about, the amount of money that we actually invested in it to get it going was not that much. They were actually fairly inexpensive projects.

Then finally, I would like to go out and treasure hunt. Let's look at the old literature. Let's look at people who might be on the fringes of science, but they don't have the right connections, like the people who I talked about earlier. They're not quite in the right place to achieve the rapid scale-up of the project. But by connecting the dots between people and topics, you know what? We could design an amazing project together.

Announcing my YouTube channel



Social theory. Geopolitics. Power. New videos every week.

Recently I launched a YouTube channel. This channel provides another medium in which to share my thoughts, as well as a place to access recordings of my talks and interviews.

New content

The first several videos dive into my thoughts on institutions, history, and modern society.

- **Silicon Valley Was Wrong: The Internet Centralized Society** It is commonly claimed that the Internet has been a decentralizing force for society, providing more power to individuals who can wield the new technology. This theme is ubiquitous within hacker culture and the cyberpunk literary genre, for example. However, today we find precisely the opposite: the Internet has, on the whole, been a *centralizing* force for society. A few large media companies have massive influence over public discourse as well as access to data about the behaviors of millions of users. While this has made individuals more transparent and more legible to large institutions at great scale, I argue that it has not made those large institutions more legible to us.
- **Why America is Not an Open Society** In this video, I explore three common models given as explanations for the success of America, and argue that they don't capture the complete picture. If these common perceptions are not true, then what more nuanced theory of history explains America's success and prosperity?
- *I. America as an open, transparent society.* Do ideas rise and fall on their own merits and strength of evidence, or is it possible to manipulate public opinion towards misinformation given enough material resources? To answer this question, I explore Edward Bernays' 1928 book Propaganda, psychiatry in communist Yugoslavia, Lysenkoism, and the (lack of) transparency of modern media institutions such as Facebook.
- *II. The American public as rational, self-interested actors.* I discuss the success of Sweden's welfare state and examples of how individuals often make economic choices that depend on trust and that reflect care for others around them, as opposed to making all choices out of pure economic self-interest.
- *III. Decisions in American governance as the output of democratic processes.* In reality, many decisions are not made by officials in elected positions, because

much political steering power is instead held by entrenched bureaucracies and civil servants.

- **Will China Out-Innovate the United States?** The United States prides itself on being a hub for world innovation and on attracting top talent from all over the world. However, China's economy is now comparable to that of the United States, and its international influence is growing to match. What forces drive this rise, and will there be consequences be for American innovation? Furthermore, what can we learn by observing the books Xi Jinping keeps on his desk?
- **How to Predict the Next Global Hub** What sociological factors have made modern Silicon Valley a hub for thinkers, innovators, and entrepreneurs? The most important factors may be unexpected, and the most expected factors may be unimportant; for example, London at its peak was crime-ridden. I explore Alexandria under the patronage of the Ptolemaic dynasty, economic opportunities in Paris during the 18th century, and the social landscapes of Los Angeles, San Francisco, and Shanghai.
- **How I Learn History** When learning history, how can we reconstruct what has truly happened? The most useful method is to ingest information from primary sources directly; these sources are not filtered through somebody else's interpretation. Do in-depth case studies, read the firsthand accounts of those who were there, and reconstruct how individuals and situations were affected by the institutions and bureaucracies around them.
- **What Is Your Theory of History?** Whether they realize it or not, everyone has their own implicit theory of history. We use our theories of history to make predictions and to decide what is important at the largest scale for our societies. An unexamined theory of history, however, can easily be inconsistent in how it reasons about the past, present, and future—and poor predictions are the result. By applying systematic thinking, you can build a theory of history that is consistent and coherent.

Talks and interviews

In addition to new content, the YouTube channel provides a location for recordings of my talks and interviews.

- **Civilization: Institutions, Knowledge, and the Future** This is a talk I gave with the Foresight Institute; I've written about it here. For the YouTube channel, I've also curated some standalone excerpts from this talk:
- **The Lycurgus Cup**: What happens when a civilization's technology becomes lost for over a thousand years? What can we learn about the economic output of the Roman Empire at its peak and before its fall? What technologies might our own civilization stand to lose? When our descendants read about our achievements, will they believe us?
- **Intellectual Dark Matter**: Physicists have inferred the existence of dark matter not by direct observation per se, but by observing the force it exerts on surrounding matter. Likewise, through observing history we can infer the existence of certain knowledge that has been developed and used by historical civilizations and which, though lost to the ages, has nonetheless shaped the trajectory of future civilizations.
- **Artificial Intelligence: Existential Hope Scenarios** This is a panel discussion with Mark Miller, Jessica Cussins, and De Kai in which I propose concrete actions towards guiding AI research to safe outcomes. I also discuss how to identify the highest risk areas of research, the feasibility of regulating software, and international cooperation.

I hope you find it interesting!

Samo Burja

Totalitarian ethical systems

This is a linkpost for http://benjaminrosshoffman.com/totalitarian-ethical-systems/

(Excerpt of another conversation with my friend Mack.)

Mack: Do you consider yourself an Effective Altruist (capital letters, aligned with at least some of the cause areas of the current *movement*, participating, etc)?

Ben: I consider myself strongly aligned with the things Effective Altruism says it's trying to do, but don't consider the movement and its methods a good way to achieve those ends, so I don't feel comfortable identifying as an EA anymore.

Consider the position of a communist who was never a Leninist, during the Brezhnev regime.

Mack: I am currently Quite Confused about suffering. Possibly my confusions have been addressed by EA or people who are also strongly aligned with the stated goals of EA and I just need to read more. I want people to *thrive* and this feels important, but I am pretty certain that "suffering" as I think the term is colloquially used is a really hard thing to evaluate, so "end suffering" might be a dead end as a goal

Ben: I think the frame in which it's important to evaluate global states using simple metrics is kind of sketchy and leads to people mistakenly thinking that they don't know what's good locally. You have a somewhat illegible but probably coherent sense that capacity and thriving are important, and that suffering matters in the context of the whole minds experiencing the suffering, not atomically

There's not actually a central decisionmaker responsible for all the actions, who has to pick a metric to add up all the goods and bads to decide which actions to prioritize. There are a lot of different decisionmakers with different capacities, who can evaluate or generate specific plans to e.g. alleviate specific kinds of suffering, and counting the number of minds affected and weighting by impact is one thing you might do to better fulfill your values. And one meta-intervention might be centralizing or decentralizing decisions.

Since you wouldn't need to do this if the info were already processed, the best you can do really is try to see (a) how different levels of centralization have worked out in terms of benefiting from economies of scale vs costs due to value-distortion in the past, and (b) whether there's a particular class of problem you care about that requires one or the other.

So, for instance, you might notice that factory farming creates a lot of pointless (from the animal's perspective) suffering that doesn't enable growth and thriving, but results from constantly thwarted intentions. This is pretty awful, and you might come up with one of many plans to avert that problem. Then you might, trying to pool resources to enact such a plan, find that other people have other plans they think are better, and try to work out some way to decide which large-scale plans to use shared resources to enact. (Assuming everyone with a large-scale plan thinks it's better than smaller-scale plans, or they'd just do their own thing)

So, one way to structure that might be hiring specialists like GiveWell / Open Phil - that's one extreme where a specialized group of plan-comparers are entrusted with

the prioritization. At the other extreme there are things like donor lotteries, where if you have X% of the funds to do something, the expected value of participating has to be at least X% of the value of funding the thing. And somewhere in the middle is some combination of direct persuasion and negotiation / trade.

Only if you go all the way to the extreme of total central planning do you really need a single totalizing metric, so to some extent proposing such a metric is proposing a totalitarian central planner, or at least a notional one like a god. This should make us at least a little worried about the proposal if it seems like the proposers are likely to be part of the decisionmaking group in the new regime. E.g. Leninism.

Mack: I'm...very cognizant of my uncertainty around what's good for other people, in part because I am often uncertain about what's good for *me.*

Ben: Yeah, it's kind of funny in the way Book II (IIRC) of Plato's Republic is funny. "I don't know what ***I*** want, so maybe I should just add up what ***everyone in the world*** wants and do that instead..."

"I don't know what a single just soul looks like, so let's figure out what an ENTIRE PERFECTLY JUST CITY looks like, and then assume a soul is just a microcosm of that."

Mack: Haven't read it, heard his Republic is a bit of a nightmare.

Ben: Well, it's a dialogue Socrates is having with some ambitious young Spartaphilic aristocrats. He points out that their desire to preserve class differences AND have good people in charge requires this totalitarian nightmare (since more narrowminded people will ALSO want the positions of disproportionate power - to be captain of the Titanic, to use a metaphor from earlier - I actually stole the ship metaphor from *Republic* - and be less distracted by questions of "how to steer the ship safely.")

He describes how even a totalitarian nightmare like this will break down in stages of corruption, and then suggests that maybe they just be happy with what they have and mostly leave other people alone.

Mack: That seems like...replacing a problem small enough for the nuance to intimidate you with one large enough that you can abstract away the nuance that would intimidate you if you acknowledged the nuance

Ben: Yes, it's not always a bad idea to try. But, like, it's one possible trick for becoming unconfused, and deciding a priori to stick with the result even if it seems kind of awful isn't usually gonna be a good move. You still gotta check that it seems right and nonperverse when applied to particular cases, using the same metrics that motivated you to want to solve the problem in the first place.

Nash equilibriums can be arbitrarily bad

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

Go hungry with Almost Free Lunches

Consider the following game, called "Almost Free Lunches" (**EDIT**: this seems to be a variant of the traveller dilemma). You name any pound-and-pence amount between £0 and £1,000,000; your opponent does likewise. Then you will both get whichever amount named was lowest.

On top of that, the person who named the highest amount must give £0.02 to the other. If you tie, no extra money changes hands.

What's the Nash equilibrium of this game? Well:

- The only Nash equilibrium of Almost Free Lunches is for both of you to name £0.00.

Proof: Suppose player A has a probability distribution p_A over possible amounts to name, and player B has a probability distribution p_B over possible amounts. Let m_A be the highest amount such that $p_A(m_A)$ is non-zero; let m_B be the same, for B. Assume that (p_A, p_B) is a Nash equilibrium.

Assume further that $m_A \geq m_B$ (if that's not the case, then just switch the labels A and B). Then either $m_A >$ £0.00 or $m_A =$ £0.00 (and hence both players select £0.00).

We'll now rule out $m_A >$ £0.00. If $m_B >$ £0.00, then player A can improve their score by replacing m_A with $m'_A = m_B -$ £0.01. To see this, assume that player B has said n_B, and player A has said m_A. If $n_B < m'_A < m_A$, then player A can say m'_A just as well as m_A - either choice gives them the same amount (namely, $n_B -$ £0.02).

There remain two other cases. If $n_B = m'_A$, then m'_A is superior to m_A, getting m'_A (rather than $m'_A -$ £0.02). And if $n_B = m_B$, then m'_A gets $m'_A +$ £0.02 = $m_B +$ £0.01,

rather than m_B (if $m_A = m_B$) or $m_B - £0.02$ (if $m_A > m_B$).

Finally, if $m_B = £0.00$, then player A gets -£0.02 unless they also say £0.00.

Hence if $m_A > £0.00$, the p_A cannot be part of a Nash Equilibrium. Thus $m_A = £0.00$ and hence the only Nash Equilibrium is at both players saying £0.00.

Pareto optimal

There are three Pareto-optimal outcomes: (£1,000,000.00, £1,000,000.00), (£1,000,000.01, £999,999.97), and (£999,999.97, £1,000,000.01). All of them are very much above the Nash Equilibrium.

Minmax and maximin

The minmax and maximin values are also both terrible, and also equal to £0.00. This is not surprising, though, as minmax and maximin implicitly assume the other players are antagonistic to you, and are trying to keep your profits low.

Arbitrary badness with two options

This shows that choosing the Nash Equilibrium can be worse than almost every other option. We can of course increase the maximal amount, and get the Nash Equilibrium to be arbitrarily worse than any reasonable solution (I would just say either £1,000,000.00 or £999,999.99, and leave it at that).

But we can also make the Nash Equilibrium arbitrarily close to the worst possible outcome, and that without even requiring more than two options for each player.

Assume that there are four ordered amounts of money/utility: $n_3 > n_2 > n_1 > n_0$. Each player can name n_2 or n_1. Then if they both name the same, they get that amount of utility. If they name different ones, then then player naming n_2 gets n_0, and the player naming n_1 gets n_3.

By the same argument as above, the only Nash equilibrium is for both to name n_1. The maximum possible amount is n_3; the maximum they can get if they both coordinate is n_2, the Nash equilibrium is n_1, and the worst option is n_0. We can set $n_1 = n_0 + \epsilon$ and $n_3 = n_2 + \epsilon$ for arbitrarily tiny $\epsilon > 0$, while setting n_2 to be larger than n_1 by some arbitrarily high amount.

So the situation is as bad as it could possibly be.

Note that this is a variant of the prisoner's dilemma with different numbers. You could describe it as "Your companion goes to a hideous jail if and only if you defect (and vice versa). Those that don't defect will also get a dust speck in their eye."

Natural Structures and Conditional Definitions

There's a sense in which definitions are arbitrary. Words are made by humans and no-one can stop me from calling red blue and blue red if I really want to. So when people ask questions like, "What is consciousness?" or "What is free-will?", it seems quite reasonable to respond, "Just pick a definition. These terms can be defined many different ways and it's completely your choice which one you choose to use".

This may appear to dissolve the question, however, I would suggest that such an answer often misunderstands what the asker is attempting. Typically the asker is concerned by more than the linguistic question, but also with attempting to understand the ontology or structure of reality. And it may be the case that this structure includes a substructure that naturally fits with our intuitions of what consciousness is or what freewill is or it may be the case, as per the standard LW view of these two cases, that such a structure doesn't exist.

What makes this especially confusing is that many people will *conditionally accept* the "it's arbitrary" answer when they are convinced that such a natural structure doesn't exist, while pointing out the natural structure otherwise. Here's an example. Let's suppose it was common knowledge that we all have souls. Then whenever someone asked about the definition of consciousness, we'd be tempted to point to the soul, just as whenever people ask about the definition of trees, we'd be tempted to talk about leaves and branches. The arguments for being able to use language arbitrarily and the fact that this isn't a perfectly well-specified definition remain. It's just that one definition suffices for 95% of cases, so we don't bring up that argument. But if instead it was common knowledge that there are no souls, it'd be much more likely they'd say that the definition is arbitrary. And by accepting answers to different questions depending on how things turn out, the intent behind the original question can easily be obscured.

Appendix

Here are some possible interpretations of, "What is X?":

- What does term X intrinsically mean? (no intrinsic meaning exists)
- What natural structure (if any) corresponds to X?
- What are some useful interpretations of the term X?
- How is the term X used in society?

These kinds of discussions tend to work better if everyone is on the same page about what is being asked.

Note: Apparently, I actually had an old post on this topic here which I'd completely forgotten about.

I've also discovered that the philosophical term is natural kinds.

More Notes on Simple Rules

Original Post by Robin: Simple Rules

Previously: Simple Rules of Law

Sarah Constantin on Twitter (if you are doomed to be on social media at all, you should follow her) offers a commentary thread of alternate explanations for the pattern pointed out in Robin's post. It contains some ideas that I didn't cover and deserve to be addressed, so this post will do that, and capture her arguments in an easier-to-find-in-the-future location. Quotes are from the thread.

> https://t.co/QKnSNtqC8r @robinhanson's explanation for why people prefer discretion to simple rules is overconfidence — everyone assumes *they'd* be the one to have special pull with decision-makers, or wants to pretend they are. "Let's play fair" is a loser's position.
>
> — Sarah Constantin (@s_r_constantin) May 20, 2019

> 1.) Low trust: nobody believes a "fair" rule would actually be applied fairly, they aren't considering the possibility of a genuinely impartial rule (and sometimes they're right)

In this scenario, people would prefer rule of law via a simple and fair rule to arbitrary decision making. But when someone proposes such a rule, or tells them they will be bound by such a rule, they do not believe it. They think that this is code for them having to abide by the rule when it suits those in power, only to have those in power ignore the rule when they would rather something else happen. So you need to obey the rules or get blamed for it, and others don't. Or, alternatively, the rules are an additional thing to be scapegoated for when the political winds call for it, without taking away any of the other methods. If you obey the rules they get you for ignoring what really matters. If you focus on what really matters they get you for not obeying the rules.

This happens often enough that the rules are now just another way to get gas-lit. Following the rules truly becomes a loser's position, in every sense.

I'd consider this a 'good' objection, similar to the Goodhart's Law objections. Simple rules, and rule of law, require high enough trust and willingness to uphold those rules, or they become another source of complexity and a tool of power. If you don't have the trust, it must be established slowly over time, or you need to structure things such that this trust emerges.

The word trust is overloaded, so there are things often called 'trust' that are *directly opposed* to the kind of trust that enables simple rules. It would be better if those things used a different word. Words need to mean things and this is one place where the lack of that makes it hard to communicate or even think clearly.

Side note: This is mostly a topic for another day that I really hope isn't what the comments talk about, but feels worth motioning at here: One answer to the question "What is blockchain?" is that blockchain enables trustless systems. Another

perspective is that *blockchain is rule of law.* You have faith that the system will follow the rules of the system, because humans don't have a reasonable way to prevent that. The weirdness comes from many of the actors in the system *being even less trustworthy than usual* because you created a system that lets you work around lack of trust and lacks the commonly used tools to punish bad behaviors. You're relying on the rule of law, which depends on choosing good rules and correctly and securely implementing them. Where people choose and correctly implement the right incentives and structures, great things can happen. Where they have other goals or choose poorly… not so much. So you simultaneously get things you can trust a lot, and a lot of lying and fraud along side that, often built directly on top of and relying directly on the trustworthy things.

> 2.) Ignorance or lack of intelligence: the idea of fair, impartial rules is a bit abstract, and has to be taught, and not everybody gets taught and not everybody copes well with abstraction.

I have a five year old son. Based on my observations of him and other children, I do not think that the idea of fair, impartial rules has to be taught. I think it is a strong instinct that things work this way. Perhaps that's because my kid is my kid, but from what I can tell younger kids (e.g. ages 5-8) are very big on what the rule is and what is fair. Monkey see, monkey do, and monkey enforce local norms. The kids will still *lie* and they'll still *break the rules* when it suits them, but they totally get it. Then, as they get older, they learn more subtle ways to break and twist the rules for advantage, and play more complex games.

I *do* think that the idea that following simple rules and having rule of law can get better overall results needs to be taught, trust needs to be learned and maintained and built, and that there's subtle stuff going on that's hard to grasp. I mostly want to blame this on bad culture and failure to teach, rather than on lack of intelligence. Most people couldn't figure this stuff out on their own, which makes it our collective responsibility to give them the tools to get there. Abstraction is by default hard, and we need concrete examples and stories and traditions that resonate – we need culture. The problem is that the powerful, and powerful natural forces, are actively fighting against us, as discussed in the previous post. And as Sarah notes next.

> (I sometimes think that if civics isn't taught in schools people will eventually grow up without actually grokking the idea of "checks on power" being a good thing *independent* of who's in power.)

Quite so. And of course, most schools do not effectively teach civics. I don't think most people get that checks on power are good, only checks on particular powerful people and things. The authoritarian instinct runs deep. Those with power would prefer people not learn that.

> 3.) Power. Often we have a discretionary rather than rule-based system not because *most* people like it that way, but because the *powerful* people like it that way. (as **@TheZvi** also said.) It's TurboTax lobbyists, not regular people, who prevent automatic tax filing.

Full agreement here. There are many things most people don't support. Because most things are not up to most people.

> 4.) Price discrimination. Often, you can get a better deal if you ask for a favor (or bargain) face to face than if you follow procedure. The average person isn't

> overconfidently estimating their charm: they're *correct* that askers do better than nonaskers on average.

Askers do better than non-askers unless asking is explicitly costly or punished. Otherwise, you sometimes get a yes and profit, and other times you get a no, and break even. And asking is usually much cheaper than it looks. We instinctively get nervous about asking, think it is risky, when usually it isn't.

Asking more often is a proven winning life strategy.

Price discrimination maximizes profits so everyone would like to engage in price discrimination, but no one wants to be subject to it. Again, we get stories of theft and power. We also get stories of forbidden considerations, and of letting ourselves consider all the data and avoid Goodhart's Law issues. Asking allows us to say no without explaining our reasons. Rules-based price discrimination seems universally hated, whereas discretionary price discrimination is mostly seen as good. I think this plays strongly into baselines and rewards versus punishments, which we'll get to at #6.

Asking is more of a method of complexity than it is an explanation for it.

Most people, from what I can tell, *strongly dislike* having to ask for things and having to haggle and navigate uncertainty. A few people like the advantage they get from being better at it, again a story largely about power and theft.

> 5.) "Copenhagen interpretation of ethics" = condemnation of intentional but not unintentional harm. This makes some sense as a legal standard, but it's crazy when you expand it to policy, as many do.
>
> Most people prefer policies with large, harmful unintended consequences over policies which explicitly admit to causing some, smaller harms. This seems like a result of confusing the question of "would this be a good world to live in?" with "should these people be punished?"

The Copenhagen Interpretation of Ethics runs even deeper than that, and was one of my five core stories. Even in this less deep form, it's still key and highly toxic. Somehow, "spend money that will require higher taxes" or "tax everyone" or the general "start with a lower baseline to reserve resources for special treatment" don't seem to trigger people's notions of intentional harm. They should. It's intentional harm, *much more directly* than many things that are objected to as intentional harm.

> 6.) There's a weird thing where justice/rationality/impersonal principle is coded as "mean" while making exceptions is coded as "nice." **A "judgmental" person is one who makes *harsh* judgments** — even though judgments can be good as well as bad.
>
> This may just be loss aversion or pessimistic bias: the fear of being punished for our failings is more salient than the hope of being rewarded for our merits.

Emphasis mine. I don't think this is loss aversion.

When people say someone is judgmental, or is judging, they're usually talking only about *negative* judgment. Rarely about positive judgment. They are the same and imply each other, but people don't see it that way. The mean thing is bad. The nice thing is good.

Which makes is super important to code actions as nice rather than mean. Actions are already on thin ice. Any system of action, whether by rule or by discretion, needs to do its best to be seen as taking nice-coded action and avoiding mean-coded action. There's also the practical problem that confiscating things is not typically something one can simply do, whereas bestowing them is allowed.

If you're dividing resources, that means you want to set the inaction baseline distribution as low as possible. Then you can make 'exceptions' to the baseline, and pay out rewards, be seen as nice, and enjoy the power from selecting the distribution of the resources you withheld or confiscated.

Screwing over the baseline scenario is not merely an incidental effect. It is a goal. It is necessary.

That (give or take a comment thread) should wrap things up. Ideally these thoughts can then be distilled into posts that are easier to make evergreen so we can build upon them better.

Physical linguistics

This is really my attempt at approaching eliminative materialism, and probably reading Paul Churchland or Daniel Dennett's papers would be better for you to get the point. I'm just writing to organize my thoughts.

Background

There are three big problems in science: universe, life, and consciousness. There is a good theory of the universe on the macro and micro scale, and the problem of its origin. They are not the final word, but we have a good sense of any future updated theories would be like: mechanistic, mathematical, probably using real, complex, and discrete numbers.

A theory of life is still in the works, though there are encouraging attempts. The physical construction of life and the *descriptive* theory of life is now complete except in the details. We know that it would be something made of evolution, thermodynamics, chemistry, and of course, mathematics. The *engineering* theory of life is still greatly missing. We do not know how to create life, at most we can fork the genetic code and do little modifications and mixings. We don't even know if a robot is alive.

A theory of consciousness is in an even earlier stage. There are some basic studies of the description of consciousness, and there are dozens of hazy philosophical theories that needs to be made quantitative using future data.

Consciousness-free

One problem with consciousness is its paradoxical qualities, creating questions that seem to both be compelling and deformed:

- "Why am I me instead of someone else?"
- "If Pinkie is copied, which one is the real Pinkie?"
- "Is the feeling of blue same for everyone?"
- "How does one freely choose?"

Now compare them with analogous questions from universe and life:

- "Why is this rock this rock instead of that rock?"
- "If this book is copied, which one is the real book?"
- "Is this website the same website on every computer?"
- "How does a slime mold decide which way to go?"

The analogous questions lose their mystery and becomes mundane, confused, or fascinating but also scientifically analyzable.

Possibly the problem is with the understanding of conscious itself, which is too confused. I propose to remove consciousness from explanations of life behavior (human or not) as much as possible. If it can be fully removed, then the problem of consciousness is solved. If it can't be fully removed, then it concentrates the effort for solution.

A sketch

As a sketch of how such a removal might be done, consider a fully physical explanation of how humans talk, which is currently infested with consciousness. The standard account is that there is a consciousness that feels something, then formulates that into words and sentences, then expresses them. Unconscious speaking is considered nonsense, meaningless, noise. This doesn't have to be.

The Heptapods

The Heptapods from Story of Your Life (Ted Chiang, 1998) are an example of a "free-will-free" form of life. Their language has determinism baked into it, just as human language has free will baked into it.

What kind of universe could produce two kinds of life such that one is deterministic in language, but the other is free in language? And in such proximity too, such that they can actually meet each other and share the same physical space and physical laws?

To answer such questions in a physics way, one would use a physics of language. What is a language according to a physicist?

Physical linguistics

What is a deterministic language, and what is a free-will language? How would a description of free-will emerge in a deterministic system such as our universe? And most importantly, how does a *universal* language, a symbolic system that can models the physical world that it is in, emerge in a deterministic world?

This is analogous to the problem of zombie language: I once read that in a world with only philosophical zombie humans, human languages would probably have not evolved to talk about consciousness and inner experiences because there is no such nonexistent thing. This argument is dumb, since human languages already talk about many nonexistent things, but it points at an interesting question: how would a deterministic system evolve a language that talks about things happening in it?

Physical self-referential science

In the same spirit, what kind of deterministic universe would have little bundles of matter inside of it that behaves roughly the same as some other patches of this universe? We call these little bundles of matter "computers running physical simulations", or "a human brain thinking about science", or maybe even "a lion brain thinking about which way an antelope is probably going to go next".

If such explanations can be done in detail, that would be a self-reference in physics: a physical system (our universe) containing a substantial description of itself (the explanation), as well as an account for why it is likely for the description to exist in the first place (the explanation about why a physical world is likely to contain its own description).

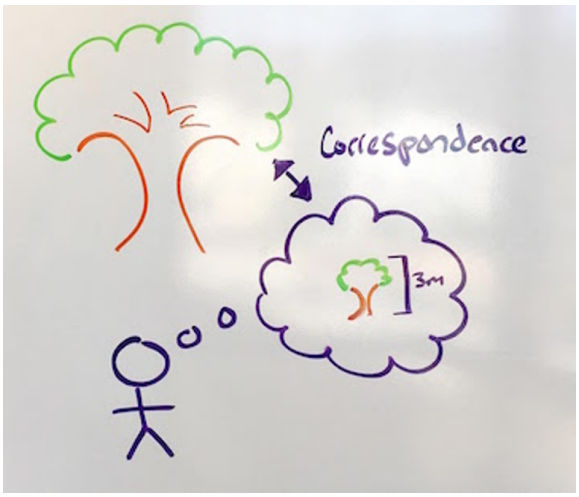
Correspondence visualizations for different interpretations of "probability"

(*Written for Arbital in 2016.*)

Recall that there are three common interpretations of what it means to say that a coin has a 50% probability of landing heads:

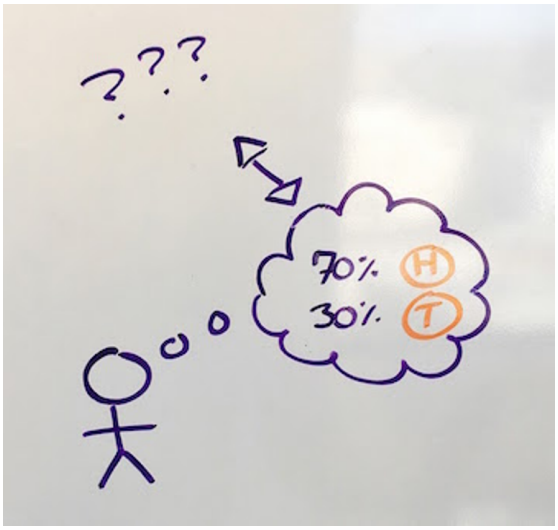
- **The propensity interpretation:** Some probabilities are just out there in the world. It's a brute fact about coins that they come up heads half the time; we'll call this the coin's physical "propensity towards heads." When we say the coin has a 50% probability of being heads, we're talking directly about this propensity.
- **The frequentist interpretation:** When we say the coin has a 50% probability of being heads after this flip, we mean that there's a class of events similar to this coin flip, and across that class, coins come up heads about half the time. That is, the *frequency* of the coin coming up heads is 50% inside the event class (which might be "all other times this particular coin has been tossed" or "all times that a similar coin has been tossed" etc).
- **The subjective interpretation:** Uncertainty is in the mind, not the environment. If I flip a coin and slap it against my wrist, it's already landed either heads or tails. The fact that I don't know whether it landed heads or tails is a fact about me, not a fact about the coin. The claim "I think this coin is heads with probability 50%" is an *expression of my own ignorance,* which means that I'd bet at 1 : 1 odds (or better) that the coin came up heads.

One way to visualize the difference between these approaches is by visualizing what they say about when a model of the world should count as a good model. If a person's model of the world is definite, then it's easy enough to tell whether or not their model is good or bad: We just check what it says against the facts. For example, if a person's model of the world says "the tree is 3m tall", then this model is correct if (and only if) the tree is 3 meters tall.

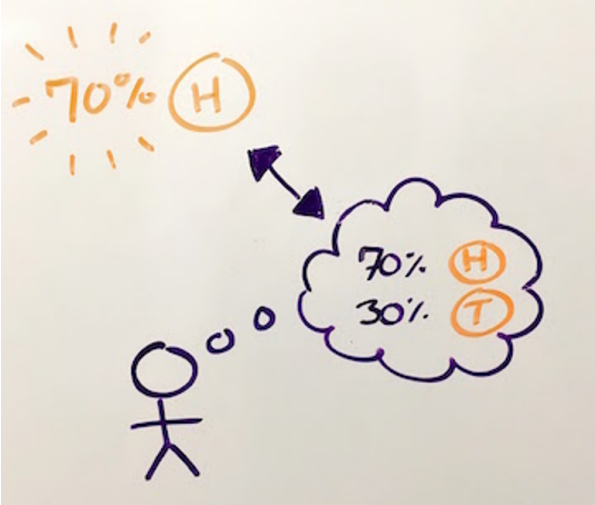


Definite claims in the model are called "true" when they correspond to reality, and "false" when they don't. If you want to navigate using a map, you had better ensure that the lines drawn on the map correspond to the territory.

But how do you draw a correspondence between a map and a territory when the map is probabilistic? If your model says that a biased coin has a 70% chance of coming up heads, what's the correspondence between your model and reality? If the coin is actually heads, was the model's claim true? 70% true? What would that mean?



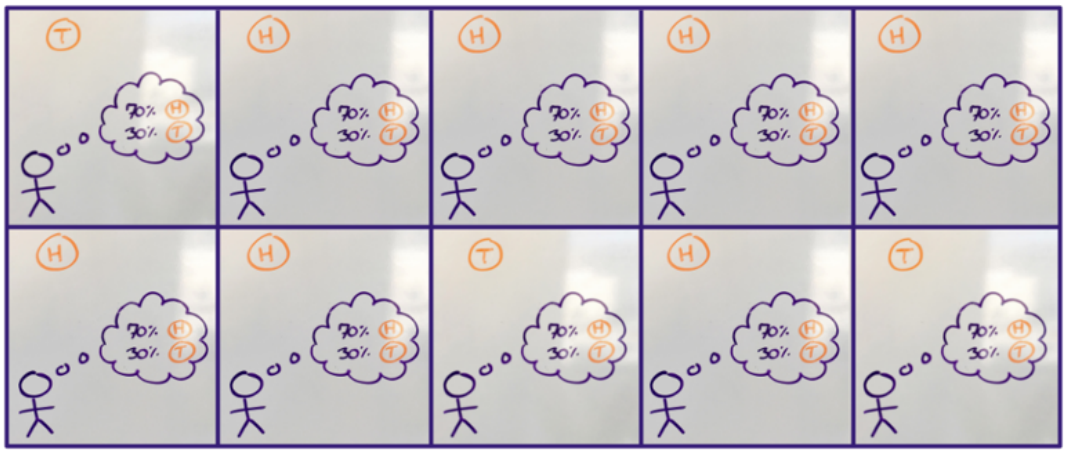
The advocate of **propensity** theory says that it's just a brute fact about the world that the world contains ontologically basic uncertainty. A model which says the coin is 70% likely to land heads is true if and only the actual physical propensity of the coin is 0.7 in favor of heads.



This interpretation is useful when the laws of physics *do* say that there are multiple different observations you may make next (with different likelihoods), as is sometimes the case (e.g., in quantum physics). However, when the event is deterministic — e.g., when it's a coin that has been tossed and slapped down and is already either heads or tails — then this view is largely regarded as foolish, and an example of the mind projection fallacy: The coin is just a coin, and has no special internal structure (nor special physical status) that makes it *fundamentally* contain a little 0.7 somewhere inside it. It's already either heads or tails, and while it may *feel* like the coin is fundamentally uncertain, that's a feature of your brain, not a feature of the coin.

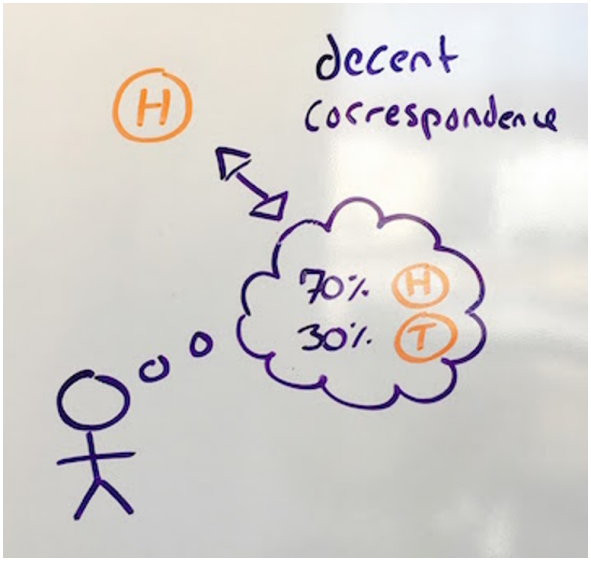
How, then, should we draw a correspondence between a probabilistic map and a deterministic territory (in which the coin is already definitely either heads or tails?)

A **frequentist** draws a correspondence between a single probability-statement in the model, and multiple events in reality. If the map says "that coin over there is 70% likely to be heads", and the actual territory contains 10 places where 10 maps say something similar, and in 7 of those 10 cases the coin is heads, then a frequentist says that the claim is true.



Thus, the frequentist preserves black-and-white correspondence: The model is either right or wrong, the 70% claim is either true or false. When the map says "That coin is 30% likely to be tails," that (according to a frequentist) means "look at all the cases similar to this case where my map says the coin is 30% likely to be tails; across all those places in the territory, 3/10ths of them have a tails-coin in them." That claim is definitive, given the set of "similar cases."

By contrast, a **subjectivist** generalizes the idea of "correctness" to allow for shades of gray. They say, "My uncertainty about the coin is a fact about *me,* not a fact about the coin; I don't need to point to other 'similar cases' in order to express uncertainty about *this* case. I know that the world right in front of me is either a heads-world or a tails-world, and I have a probability distribution that puts 70% probability on heads." They then draw a correspondence between their probability distribution and the world in front of them, and declare that the more probability their model assigns to the correct answer, the better their model is.



If the world *is* a heads-world, and the probabilistic map assigned 70% probability to "heads," then the subjectivist calls that map "70% accurate." If, across all cases where their map says something has 70% probability, the territory is actually that way 7/10ths of the time, then the Bayesian calls the map "well calibrated". They then seek methods to make their maps more accurate, and better calibrated. They don't see a need to interpret probabilistic maps as making definitive claims; they're happy to interpret them as making estimations that can be graded on a sliding scale of accuracy.

Debate

In short, the frequentist interpretation tries to find a way to say the model is definitively "true" or "false" (by identifying a collection of similar events), whereas the subjectivist interpretation extends the notion of "correctness" to allow for shades of gray.

Frequentists sometimes object to the subjectivist interpretation, saying that frequentist correspondence is the only type that has any hope of being truly objective. Under Bayesian correspondence, who can say whether the map should say 70% or 75%, given that the probabilistic claim is not objectively true or false either way? They claim that these subjective assessments of "partial accuracy" may be intuitively satisfying, but they have no place in science. Scientific reports ought to be restricted to frequentist statements, which are definitively either true or false, in order to increase the objectivity of science.

Subjectivists reply that the frequentist approach is hardly objective, as it depends entirely on the choice of "similar cases". In practice, people can (and do!) abuse frequentist statistics by choosing the class of similar cases that makes their result look as impressive as possible (a technique known as "p-hacking"). Furthermore, the manipulation of subjective probabilities is subject to the iron laws of probability theory (which are the only way to avoid inconsistencies and pathologies when managing your uncertainty about the world), so it's not like subjective probabilities are the wild west or something. Also, science has things to say about situations even when there isn't a huge class of objective frequencies we can observe, and science should let us collect and analyze evidence even then.

For more on this debate, see Likelihood functions, p-values, and the replication crisis.

Space colonization: what can we definitely do and how do we know that?

Arguments for the value of the long-term future tend to make the assumption that we will colonize space. What can we definitely accomplish in terms of space colonization? Why think that we can definitely do those things?

The FHI paper, Eternity in Six Hours, is very optimistic about what can be done:

> In this paper, we extend the Fermi paradox to not only life in this galaxy, but to other galaxies as well. We do this by demonstrating that traveling between galaxies – indeed even launching a colonisation project for the entire reachable universe – is a relatively simple task for a star-spanning civilization, requiring modest amounts of energy and resources. We start by demonstrating that humanity itself could likely accomplish such a colonisation project in the foreseeable future, should we want to, and then demonstrate that there are millions of galaxies that could have reached us by now, using similar methods.

Is this paper reasonable? Which parts of its assertions are most likely to be mistaken?

This question was inspired by a conversation with Nick Beckstead.

What is required to run a psychology study?

I periodically find myself wishing someone had run an experiment on a particular topic. But they haven't.

Often, it seems like there's a relatively easy experiment you can run, which would give me some evidence about it. Maybe it wouldn't be perfect evidence, but it would be better than me making stuff up from my armchair.

The two clusters of things-I-can-imagine-being-quite-hard are:

- Legal requirements
- Actually doing the science right (avoiding sampling biases, actually testing the right variable, etc)

Legal Stuff

Scott's IRB Nightmare suggests that the legal requirements can be quite awful. I'm unclear on what those requirements are if I don't want to publish anywhere, don't plan to interface with any hospital bureaucracies, I just want to ask a bunch of people to try something and see how it goes (and maybe make a LessWrong blogpost so people on LessWrong can know too)

Am I allowed to just go out and ask a whole bunch of people stuff? If I want to know things about 8 year olds or 16 year olds, if their parents sign a form and I'm not doing anything especially weird or traumatic, would that be fine or would terrible things happen to me? (Could I replicate the Marshmallow Test?)

Doing Science Good Enough

I have a sense that there's a lot of ways to deceive yourself if you have a pet psychology theory. But... I dunno it also seems like LessWrong collectively should be pretty good at this. The thing I might want to do (or get others to do sometimes) is post a plan for a study here, and get critiqued until it seems like an actually good plan, do the plan, write up the results.

Particular Things I was interested in:

(not meant to be exhaustive, important, or particularly achievable plans, just the things that generated this question)

- How do most people relate to their job? (There's a survey study that asks if people consider it a job, a career, or a calling, but the study had a very narrow range of participants – people who worked at one particular place, and I'm curious how a wider variety of people would respond)
- How do most people relate to ambition? (where by ambition I mean "form plans to create or change things that will impact large numbers of people.")
- How many 8 year olds can learn to program? Can they implement FizzBuzz? How many 16 year olds? How many 30 year olds? I've heard that people either have-or-don't-have a "programmer trait" that's hard to learn, and I'm not sure how much of that is true at all, and if so if it's more about nature or nurture.

345M version GPT-2 released

This is a linkpost for https://openai.com/blog/better-language-models/#update

OpenAI has released a larger GPT-2 model for public testing. They've also released the two larger models to select groups for experimenting.

What makes a scientific fact 'ripe for discovery'?

The existence of multiple discovery seems to suggest that there are certain factors that make scientific facts ready to be discovered. What are these, factors, and how could one measure them?

Evidence for Connection Theory

This is a linkpost for https://www.scribd.com/document/219774356/Evidence-for-Connection-Theory#fullscreen

Connection Theory (CT) is the original philosophy underpinning Leverage Research, a research think tank focused that has worked with the effective altruism movement, and the rationality community, in the past on community-building, and existential risk reduction. CT was developed by Leverage's executive director, Geoff Anders. Since there are few if any other publicly available online resources for understanding or evaluating CT, I thought I would share this document.

And the AI would have got away with it too, if...

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

Paul Christiano presented some low-key AI catastrophe scenarios; in response, Robin Hanson argued that Paul's scenarios were not consistent with the "large (mostly economic) literature on agency failures".

He concluded with:

> For concreteness, imagine a twelve year old rich kid, perhaps a king or queen, seeking agents to help manage their wealth or kingdom. It is far from obvious that this child is on average worse off when they choose a smarter more capable agent, or when the overall pool of agents from which they can choose becomes smarter and more capable. And its even less obvious that the kid becomes maximally worse off as their agents get maximally smart and capable. In fact, I suspect the opposite.

Thinking on that example, my mind went to Edward the Vth of England (one of the "Princes in the Tower"), deposed then likely killed by his "protector" Richard III. Or of the Guangxu Emperor of China, put under house arrest by the Regent Empress Dowager Cixi. Or maybe the ten year-old Athitayawong, king of Ayutthaya, deposed by his main administrator after only 36 days of reign. More examples can be dug out from some of Wikipedia's list of rulers deposed as children.

We have no reason to restrict to child-monarchs - so many Emperors, Kings, and Tsars have been deposed by their advisers or "agents". So yes, there are many cases where agency fails catastrophically for the principal and where having a smarter or more rational agent was a disastrous move.

By restricting attention to agency problems in economics, rather than in politics, Robin restricts attention to situations where institutions are strong and behaviour is punished if it gets too egregious. Though even today, there is plenty of betrayal by "agents" in politics, even if the results are less lethal than in times gone by. In economics, too, we have fraudulent investors, some of which escape punishment. Agents betray their principals to the utmost - when they can get away with it.

So Robin's argument is entirely dependent on the assumption that institutions or rivals will prevent AIs from being able to abuse their agency power. Absent that assumption, most of the "large (mostly economic) literature on agency failures" becomes irrelevant.

So, would institutions be able to detect and punish abuses by future powerful AI agents? I'd argue we can't count on it, but it's a question that needs its own exploration, and is very different from what Robin's economic point seemed to be.

A shift in arguments for AI risk

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://fragile-credences.github.io/prioritising-ai/

The linked post is work done by Tom Adamczewski while at FHI. I think this sort of expository and analytic work is very valuable, so I'm cross-posting it here (with his permission). Below is an extended summary; for the full document, see his linked blog post.

Many people now work on ensuring that advanced AI has beneficial consequences. But members of this community have made several quite different arguments for prioritising AI.

Early arguments, and in particular *Superintelligence*, identified the "alignment problem" as the key source of AI risk. In addition, the book relies on the assumption that superintelligent AI is likely to emerge through a discontinuous jump in the capabilities of an AI system, rather than through gradual progress. This assumption is crucial to the argument that a single AI system could gain a "decisive strategic advantage", that the alignment problem cannot be solved through trial and error, and that there is likely to be a "treacherous turn". Hence, the discontinuity assumption underlies the book's conclusion that existential catastrophe is a likely outcome.

The argument in *Superintelligence* combines three features: (i) a focus on the alignment problem, (ii) the discontinuity assumption, and (iii) the resulting conclusion that an existential catastrophe is likely.

Arguments that abandon some of these features have recently become prominent. They also generally tend to have been made in less detail than the early arguments.

One line of argument, promoted by Paul Christiano and Katja Grace, drops the discontinuity assumption, but continues to view the alignment problem as the source of AI risk. Even under more gradual scenarios, they argue that, unless we solve the alignment problem before advanced AIs are widely deployed in the economy, these AIs will cause human values to eventually fade from prominence. They appear to be agonistic about whether these harms would warrant the label "existential risk".

Moreover, others have proposed AI risks that are unrelated to the alignment problem. I discuss three of these: (i) the risk that AI might be misused, (ii) that it could make war between great powers more likely, and (iii) that it might lead to value erosion from competition. These arguments don't crucially rely on a discontinuity, and the risks are rarely existential in scale.

It's not always clear which of the arguments actually motivates members of the beneficial AI community. It would be useful to clarify which of these arguments (or yet other arguments) are crucial for which people. This could help with evaluating the strength of the case for prioritising AI, deciding which strategies to pursue within AI, and avoiding costly misunderstanding with sympathetic outsiders or sceptics.

Von Neumann’s critique of automata theory and logic in computer science

Quote from The General and Logical Theory of Automata. Corrected some typos using this version. H/T Hacker News.

> There exists today a very elaborate system of formal logic, and, specifically, of logic as applied to mathematics. This is a discipline with many good sides, but also with certain serious weaknesses. This is not the occasion to enlarge upon the good sides, which I have certainly no intention to belittle. About the inadequacies, however, this may be said: Everybody who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or of the complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best-elaborated part of mathematics. **Thus formal logic is, by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of the mathematical terrain, into combinatorics**.
>
> The theory of automata, of the digital, all-or-none type, as discussed up to now, is certainly a chapter in formal logic. It would, therefore, seem that it will have to share this unattractive property of formal logic. It will have to be, from the mathematical point of view, combinatorial rather than analytical.
>
> *Probable Characteristics of Such a Theory.* Now it seems to me that this will in fact not be the case. In studying the functioning of automata, it is clearly necessary to pay attention to a circumstance which has never before made its appearance in formal logic.
>
> Throughout all modern logic, the only thing that is important is whether a result can be achieved in a finite number of elementary steps or not. The size of the number of steps which are required, on the other hand, is hardly ever a concern of formal logic. Any finite sequence of correct steps is, as a matter of principle, as good as any other. It is a matter of no consequence whether the number is small or large, or even so large that it couldn’t possibly be carried out in a lifetime, or in the presumptive lifetime of the stellar universe as we know it. In dealing with automata, this statement must he significantly modified. In the case of an automaton the thing which matters is not only whether it can reach a certain result in a finite number of steps at all but also how many such steps are needed. There are two reasons. First, automata are constructed in order to reach certain results in certain pre-assigned durations, or at least in pre-assigned orders of magnitude of duration. Second, the componentry employed has in every individual operation a small but nevertheless non-zero probability of failing. In a sufficiently long chain of operations the cumulative effect of these individual probabilities of failure may (if unchecked) reach the order of magnitude of unity-at which point it produces, in effect, complete unreliability. The probability levels which are involved here are very low, but still not too far removed from the domain of ordinary technological experience. It is not difficult to estimate that a

high-speed computing machine, dealing with a typical problem, may have to perform as much as 10^12 individual operations. The probability of error on an individual operation which can be tolerated must, therefore, be small compared to 10^-12. I might mention that an electromechanical relay (a telephone relay) is at present considered acceptable if its probability of failure on an individual operation is of the order 10^-8. It is considered excellent if this order of probability is 10^-9 Thus the reliabilities required in a high-speed computing machine are higher, but not prohibitively higher, than those that constitute sound practice in certain existing industrial fields. The actually obtainable reliabilities are, however, not likely to leave a very wide margin against the minimum requirements just mentioned. An exhaustive study and a nontrivial theory will, therefore, certainly be called for.

Thus the logic of automata will differ from the present system of formal logic in two relevant respects.

1. The actual length of “chains of reasoning,” that is, of the chains of operations, will have to be considered.

2. The operations of logic (syllogisms, conjunctions, disjunctions, negations, etc., that is, in the terminology that is customary for automata, various forms of gating, coincidence, anti-coincidence, blocking, etc., actions) will all have to be treated by procedures which allow exceptions (malfunctions) with low but non-zero probabilities. All of this will lead to theories which are much less rigidly of an all-or-none nature than past and present formal logic. They will be of a much less combinatorial, and much more analytical, character.

Kevin Simler's "Going Critical"

This is a linkpost for https://www.meltingasphalt.com/interactive/going-critical/

An interactive blogpost by Kevin Simler on network dynamics, with a final section on academia and intellectual progress. I generally think careful exploration of small-scale simulations like this can help quite well with understanding difficult topics, and this post seems like a quite good execution of that approach.

Also some interesting comments on intellectual progress and academia (though I recommend reading the whole post):

> For years I've been fairly dismissive of academia. A short stint as a PhD student left a bad taste in my mouth. But now, when I step back and think about it (and abstract away all my personal issues), I have to conclude that academia is still *extremely* important.
>
> Academic social networks (e.g., scientific research communities) are some of the most refined and valuable structures our civilization has produced. Nowhere have we amassed a greater concentration of specialists focused full-time on knowledge production. Nowhere have people developed a greater ability to understand and critique each other's ideas. This is the beating heart of progress. It's in these networks that the fire of the Enlightenment burns hottest.
>
> But we can't take progress for granted. If the reproducibility crisis has taught us anything, it's that science can have systemic problems. And one way to look at those problems is network degradation.
>
> Suppose we distinguish two ways of practicing science: *Real Science* vs. *careerist science*. Real Science is whatever habits and practices reliably produce knowledge. It's motivated by curiosity and characterized by honesty. (Feynman: "I just have to understand the world, you see.") Careerist science, in contrast, is motivated by professional ambition, and characterized by playing politics and taking scientific shortcuts. It may look and act like science, but it *doesn't* produce reliable knowledge.
>
> (Yes this is an exaggerated dichotomy. It's a thought exercise. Bear with me.)
>
> Point is, when careerists take up space in a Real Science research community, they gum up the works. They angle to promote themselves while the rest of the community is trying to learn and share what's true. Instead of striving for clarity, they complicate and obfuscate in order to sound more impressive. They engage in (what Harry Frankfurt might call) scientific bullshit. And consequently, we might model them as dead nodes, immune to the good-faith information exchanges necessary for the growth of knowledge:

Episode 3 of Tsuyoku Naritai! (the 'becoming stronger podcast): Nike Timers

Latest episode is up! In this episode, we learn rationality from Shia LeBeouf.

https://anchor.fm/tsuyokunaritai/episodes/Episode-3---Nike-Timers-e3trl2

http://bit.ly/2GZnB0E

"One Man's Modus Ponens Is Another Man's Modus Tollens"

This is a linkpost for https://www.gwern.net/Modus

Eight Books To Read

This article was originally published on SamoBurja.com. You can [access the original here.]

A few years ago, I was asked by a friend what news sources they should follow to understand the Syrian Civil War. I replied they shouldn't follow any news at all. My recommendation instead was a six month break from Syrian news, supplemented by leisurely reading through six books on Syrian politics, economics, and culture. I pointed out they could read them on their phone just as conveniently as they could read tweets or articles. My friend was taken aback but followed the advice.

Critiques of news media are much more in vogue now than they were in 2015. People bemoan the poor factual accuracy or manifest political bias of today's media, whether that means established newspapers like The New York Times or social networks like Facebook. But there is a more fundamental problem with news: it can provide information, but isn't structured to educate you into someone who could understand this cherry-picked information. Formal education often fails to provide this vital foundation.

After six months, my friend thanked me. They said they now barely follow any news on Syria, but when they do it has gone from perplexing to understandable. The fragments of information no longer landed only as emotional bursts of excitement or anxiety, but rather helped contribute to a solid picture of the region. They asked me a more difficult question: what books should they read to understand not just Syria, but global society as a whole?

Books are incomplete instruments for instruction. They don't respond to the reader and cannot directly answer questions, and they require a strange and systematic process of study that goes beyond mere reading. In physics education, for example, one will pair up the mastery of theories with tests of solving mathematical puzzles as well as a course of practical experiments that tie those to one's senses. For the study of society, there would have to be analogues.

Further, true autodidacticism is a rare gift. To maintain motivation over a few months, learning has to be its own reward. This reward of learning must somehow be tied to understanding the world as it is, rather than pursuing theories for the sake of entertainment.

Much has happened throughout human history, and much is happening right now. Too much to ever fully catch up on. The focus should rather be on equipping someone with the theory and skills needed so they will process, absorb, and retain the information they encounter throughout their intellectual lives. This merits a methodological approach tailored to individual investigation and practical application.

The order in which one reads also matters. Important parts of certain books are unlocked by the understanding gained from another. This is obvious for disciplines like theoretical physics, but the same goes for a serious study of society.

While I have made it my core area of research, I can't claim to fully understand society. All I could do was try to think of the most efficient way to acquire a measure of competency in the areas I pursued.

So with those caveats, I gave him a list of the sequence of books I recommend reading:

1. **How to Read a Book—Mortimer Adler**

This book convinced me that while skimming was perhaps useful for mining information, it would never be a viable path to rigor. Adler advocates a disciplined and deep reading of challenging books. The book lays out a systematic method that, if followed, notably increases the skill of reading comprehension to the level of most graduate programs, improving one's ability to learn from books. You can then afford to read more slowly because you gain more information from each reading. This is a necessity for the systematic study of society. Examples of books you'll want to read in such an intellectual pursuit are primary sources for case studies, books laying out political theory, economics, as well as the other books on this list. He wrote this book in the 1950s as his attempt at an antidote to shortening attention spans. Unfortunately, in the age of social media, we need his remedy much more than he could have possibly imagined.

2. **The Republic—Plato**

A design for the creation of a new ideal society, the ultimate aim of sociological investigation. Plato's work is a nice example of both the strengths and the limits of a theory-driven approach. The book not only lays out the major research tasks needed to engineer such a society, but lays out a plan to construct it.

He introduces a decent theory of psychology, allowing for some basic predictions of how people will respond to changes in their social and material circumstances. The model of learning introduced is among the best I've found.

The quality of his models is sufficient to be worth knowing and occasionally using. His variant of the theory of cycles of social transformation, of how city states change between regimes of status- and emotion-driven regulation and their related political constitutions, remains predictive. The theories of psychology, education, and society are tied together into a theory-driven design for an elite.

The practicality of this book as a manual for such efforts is underrated. As an example, it illustrates how to dialogue under adversarial circumstances. This is useful for both for your own ability to manage information, and your ability to successfully interpret texts produced under such circumstances.

3. **History of the Peloponnesian War—Thucydides**

The ultimate primary source. Thucydides fought as a general during the Peloponnesian war and was ultimately exiled from Athens due to political machination. After the end of the war, he spent his energies and wealth to follow up on the connections, both friend and foe, he had built. He thought the role of a historian was to chronicle and competently navigate the era in which he lived to preserve its lessons for the future. Because Thucydides was a practitioner, and one who played a critical role in the events described, his account should be taken seriously.

A clear enough demonstration of excellent analysis, such that one can productively use his approach as a prototype. He successfully combines information sources such as interviews, texts, and personal experience of war and politics. This is then paired with the skilful application of theoretical constructs to analyze a concrete circumstance.

4. **Politics—Aristotle**

Classical sources credit Aristotle with 170 “constitutions”, that is, research papers describing the political structure and society of various Greek city-states. Many of these constitutions were likely written or drafted by his students. Of this extensive empirical research done in preparation for the Politics, the only constitution preserved is that of Athens. This marriage of empirical data and philosophy proves hard to beat.

Aristotle’s observations on Greek society should round out what was learned following Thucydides’ exhaustive account and help complete a basic understanding of a period of history one can then reason about. Further, an alternative frame of analysis of Greek politics allows for comparison with Thucydides’ often cynical explanations.

Aristotle critiques a number of Plato’s ideas, which should improve your understanding of the Republic, help you learn how to identify potential sociological reasoning flaws, and illustrate how to refute a sociological theory.

He demonstrates how to translate the analysis of social roles and professions into a generalized analysis of a society. Sociologists and economists have divorced the two, but they are inseparable when done well. This forms the basis of class analysis as used by later thinkers like Smith, Marx, and Veblen.

It is rare for a social scientist to examine social technology as lucidly. For example, the Aristotelian account of hierarchy and why it arises is superb. The conceptually clean distinctions between the different forms of interpersonal coordination can greatly augment one’s ability to navigate and study such patterns.

Finally, as an example of a good scientist and philosopher, he can be used to help understand scientists and philosophers in general.

5. **On War—Carl von Clausewitz**

Clausewitz was a Prussian staff officer in the wars against Napoleon and later became an influential military theorist. Good military theory is rarely spread publicly, but Clausewitz’s wife was a prominent noblewoman who published his magnum opus after his death.

His work provides a demonstration of excellent theoretical sociological methodology, especially as regards the proper use of case studies, how to tell the general from the particular, and how to tell the fundamental from the subordinate.

Clausewitz's model of how armies function provides a foundation for understanding the methodology and conclusions of Great Founder Theory as applied to the military.

This book shows how an essentially philosophical approach can be brought far enough to be practically useful.

6. **Great Founder Theory—Samo Burja**

A work in progress, but a decent introduction. This explains my current sociological paradigm.

7. **The Evolution of Civilizations—Carroll Quigley**

This book provides a good macro theory of civilization. Much of the pre-historical speculation can be skipped, but the overviews of historical civilizations provide an example of first-rate institutional analysis.

Quigley's career demonstrates an excellent piece of sociological methodology around gathering information to test your theory: he builds a theory that emphasizes the importance of elites, and subsequently goes and talks to members of the elite to test and apply the theory. Note, however, that he is not a practitioner, so his usefulness as an exemplar who tests and acts on their theory is somewhat limited.

8. **Persecution and the Art of Writing—Leo Strauss**

Thinkers can provoke social or legal penalties in all societies. An important way to avoid attack that can derail a career, intellectual project, or a life is to learn to write between the lines. Leo Strauss' work helps you learn improved text interpretation procedures by teaching you to read between the lines, representing a good upgrade on what you learned from Adler. It is a very good practice to attempt a 'Straussian' reading of a text even when there is no hidden message, since it entices to a higher level of information processing.

At this point, you can continue on your journey or recurse to reread Quigley, Plato, and Thucydides. Quigley writes obliquely and at a distance regarding Anglo-American elites. Thucydides is a political exile from Athens. While Thucydides is significantly freed from constraints and retribution, he will continue to have notable conflicts of interest and messaging agendas. Plato writes trickily for pedagogical purposes, intentionally setting challenges and puzzles for the reader, hoping the reader uses his text as an obstacle course to grow stronger.

Someone who makes it through this list, if they approach the texts with the rigor advised by Adler, will have the foundational understanding necessary for interpreting social events. They will be better equipped to do so than the vast majority of people.

The most important thing to be gained from these texts is a set of methodological tools, a way of thinking about and interpreting social events, that one can then use to generate one's own insights about society. These authors try to bridge the gaps between the practitioner, the theorist, and the empiricist. This is something a great sociologist must do. One of the most tragic flaws a historian can have is a myopic interest in events, rather than societies. The most tragic flaw of a social scientist is the ignorance of history that trivially rebuts the most beautiful statistically-derived or philosophically-derived theory of society.

Secondarily, these authors provide superb examples of what good sociology looks like, which can then be used to construct one's model of real expertise in this domain. This is critical for evaluating the host of supposed experts who claim to have an understanding of society that gives authority to their interpretations of events. Separating the wheat from the chaff is necessary for navigating the contemporary discourse without being misled.

Many others have since asked me for such lists, so I've kept it around and shared it whenever my friends or acquaintances have asked for book recommendations. Now you have it. Will you take a break from the news to read and think?

Read more from Samo Burja here.

Blame games

This is a linkpost for http://benjaminrosshoffman.com/blame-games/

In Excerpts from a larger discussion about simulacra, I worked through a well-known schema for distinguishing different relationships towards semantic reference, that are a natural result of interactions between shared-production games and expropriation games. Here, I analyze the coalition politics of such games.

The Survivor game

In zero-sum games, majoritarian decision rules (such as democracy) create an asymmetry - it's much easier to expropriate from a minority than from a majority - or, easier to transfer wealth to a majority than to a minority. Why would the majority vote for something they don't all benefit from?

A simple variant of this is the Survivor game, in which a single player is voted off the island at a time (see also the ancient Greek custom of ostracism). Since there's comparatively little advantage to being singled out for good, players will tend to want to avoid revealing information about themselves or their allies. Loudly voicing consensus opinion in ways that don't specify the implications for any person is fine because it's not informative. Anything that lets people distinguish you from the others is dangerous.

The idea of a Schelling point is that if players in a game need to converge on one location in a map, then in the absence of a strong incentive to favor one location, they will tend to converge on some obviously identifiable feature. For instance, in surveys, Thomas Schelling found that a surprisingly large number of people, if tasked with meeting someone on a specified day, in New York, with no further information, would converge on the information booth in Grand Central - and if no time was specified, they favored noon.

In a pure Survivor game, the first player to reveal their "location" loses. They become the feature everyone else converges on as an expropriation target. One natural side effect of this is coordination against any players who are narratively constrained by something other than the zero-sum game. For instance, if a widget-making group isn't under sharp performance pressure, anyone who's focused on actually making the widgets is going to have a hard time staying in lockstep with the group story, and is therefore the easiest target for expropriation and exclusion.

(Compare with Sarah Constantin's claim that group-coordination activities like dancing serve as a way to identify and exclude people who are out of sync with the whole. This stands in some tension with her more recent claim that people should exaggerate differences in order to have some social standing within a group.)

The Scapegoating game

What if you try to play the Survivor game in the real world, where there are other games going on? Now your environment is not exclusively populated with zero-sum players and strategies, which means that revealing info isn't always an unforced error.

Level 1: Fault analysis

I already mentioned that people trying to coordinate in objective reality will be narratively constrained in ways that make them easier targets for expropriation. But there's another feature of group coordination that's very exploitable in the Survivor game: fault analysis. We try to improve maps to improve productive capacity and mitigate risks external to the social game. An important part of this is revealing flaws in the current arrangement.

If you reveal a flaw, you might try to repair the defect (e.g. getting someone to change their behavior - "the squeaky wheel gets the grease") or you might just discard the flawed part (e.g. punishments for bad behavior, "the squeaky wheel gets replaced"). This is what fault analysis looks like in simulacrum level 1 - the meaning of "flaw" correspond to the anticipation that if you remove the flaw, some objective problem is eliminated or ameliorated.

Level 2: Framing

If there's any amount of zero-sum conflict going on inside the group, the fault-analysis machinery - if coupled to punishment at all - becomes a weapon in the hands of anyone willing to lie. If I want to target someone for expropriation in a zero-sum conflict, I can recruit naive level-1 players by accusing them of some objective flaw - framing them.

Consider the story of the vineyard of Naboth, from 1 Kings 21:

> *And it came to pass after these things, that Naboth the Jezreelite had a vineyard, which was in Jezreel, hard by the palace of Ahab king of Samaria. And Ahab spake unto Naboth, saying, Give me thy vineyard, that I may have it for a garden of herbs, because it is near unto my house: and I will give thee for it a better vineyard than it; or, if it seem good to thee, I will give thee the worth of it in money.*
>
> *And Naboth said to Ahab, The Lord forbid it me, that I should give the inheritance of my fathers unto thee.*
>
> *And Ahab came into his house heavy and displeased because of the word which Naboth the Jezreelite had spoken to him: for he had said, I will not give thee the inheritance of my fathers. And he laid him down upon his bed, and turned away his face, and would eat no bread.*
>
> *But Jezebel his wife came to him, and said unto him, Why is thy spirit so sad, that thou eatest no bread?*
>
> *And he said unto her, Because I spake unto Naboth the Jezreelite, and said unto him, Give me thy vineyard for money; or else, if it please thee, I will give thee another vineyard for it: and he answered, I will not give thee my vineyard.*
>
> *And Jezebel his wife said unto him, Dost thou now govern the kingdom of Israel? arise, and eat bread, and let thine heart be merry: I will give thee the vineyard of Naboth the Jezreelite.*

So she wrote letters in Ahab's name, and sealed them with his seal, and sent the letters unto the elders and to the nobles that were in his city, dwelling with Naboth. And she wrote in the letters, saying, Proclaim a fast, and set Naboth on high among the people: And set two men, sons of Belial, before him, to bear witness against him, saying, Thou didst blaspheme God and the king. And then carry him out, and stone him, that he may die.

And the men of his city, even the elders and the nobles who were the inhabitants in his city, did as Jezebel had sent unto them, and as it was written in the letters which she had sent unto them. They proclaimed a fast, and set Naboth on high among the people. And there came in two men, children of Belial, and sat before him: and the men of Belial witnessed against him, even against Naboth, in the presence of the people, saying, Naboth did blaspheme God and the king. Then they carried him forth out of the city, and stoned him with stones, that he died. Then they sent to Jezebel, saying, Naboth is stoned, and is dead.

And it came to pass, when Jezebel heard that Naboth was stoned, and was dead, that Jezebel said to Ahab, Arise, take possession of the vineyard of Naboth the Jezreelite, which he refused to give thee for money: for Naboth is not alive, but dead.

And it came to pass, when Ahab heard that Naboth was dead, that Ahab rose up to go down to the vineyard of Naboth the Jezreelite, to take possession of it.

And the word of the Lord came to Elijah the Tishbite, saying, Arise, go down to meet Ahab king of Israel, which is in Samaria: behold, he is in the vineyard of Naboth, whither he is gone down to possess it. And thou shalt speak unto him, saying, Thus saith the Lord, Hast thou killed, and also taken possession? And thou shalt speak unto him, saying, Thus saith the Lord, In the place where dogs licked the blood of Naboth shall dogs lick thy blood, even thine.

And Ahab said to Elijah, Hast thou found me, O mine enemy?

And he answered, I have found thee: because thou hast sold thyself to work evil in the sight of the Lord. Behold, I will bring evil upon thee, and will take away thy posterity, and will cut off from Ahab him that pisseth against the wall, and him that is shut up and left in Israel, And will make thine house like the house of Jeroboam the son of Nebat, and like the house of Baasha the son of Ahijah, for the provocation wherewith thou hast provoked me to anger, and made Israel to sin.
[...]

And it came to pass, when Ahab heard those words, that he rent his clothes, and put sackcloth upon his flesh, and fasted, and lay in sackcloth, and went softly.

King Ahab, a level-1 player, sees no way to acquire his neighbor's vineyard lawfully. But his foreign queen Jezebel, used to higher simulacrum level royal politics, sees no impediment to simply framing Naboth, a simulacrum level 2 tactic.

Elijah sees this as an existential threat, flips out, and yells at the king that he deserves the death penalty for this, since by going along with this he's raised the simulacrum level of his kingdom, making object-level coordination harder in a way that can, if it goes too far, become irreversible. Ahab, still a level 1 player, accepts the validity of Elijah's critique and tries to learn his lesson.

When the Survivor game is coupled to fault-analysis in this way, it becomes the Scapegoat game. If the simulacrum level 1 players are naive about this, a minority of zero-sum players can quickly acquire an advantage, since they're working harder to avoid becoming expropriation targets.

Level 3: Prosecutorial discretion

When enough players are mainly using fault-analysis to play the Scapegoat game instead of to fix things, the penal code can be redefined so that nearly everyone is technically guilty of some serious crime, and prosecutorial discretion is required. Then, you don't even need to lie to target someone (thus opening yourself up to expropriation for the crime of lying) - since everyone's guilty, actually being guilty of a crime doesn't single you out anymore. The crimes that get punished are the ones where the governing majority sees a shared interest in expropriating from someone. This is simulacrum level 3, where there's no underlying consistent mapping of crimes to punishments that would be good if enforced, just a standardized list of approved attacks.

Consider the case of Martin Shkreli, who everyone hated because of some perfectly legal price gouging (not morally innocent or sympathetic like Naboth, but not actually criminal), and was consequently prosecuted for the common and totally unrelated crime of securities fraud. There's not really a norm against securities fraud in the sense of effectual coordination to prevent it from happening, there's just a norm that it's a valid accusation. It's increasingly expensive to be innocent.

(But Martin Shkreli was a bad guy and deserved prison? Whatever. Once we're arguing about that instead of trying to criminalize the behavior we actually object to, we've abandoned the pretense that the penal code is a serious attempt to represent which behavior we intend to punish.)

Completely fictitious crimes like witchcraft are a natural outgrowth of this, provided there's a mechanism for confirming that some such claims are true and therefore that the target should be punished. At the limit, we start to see fully general fault-assignment stories, such as such as the Original Sin of Adam and Eve, for which humanity was punished with babies, crops, and the ability to kill snakes, or St. Andreas's Fault, for which Californians are punished with earthquakes.

Level 4: Shoot the messenger

Finally, at simulacrum level 4, people stop tracking the objective meaning of the law even locally, and it collapses to the pure Survivor game again. Prosocial behavior like revealing information about other people's crimes (e.g. Edward Snowden and Reality Winner, but also Frank Serpico) can be enough.

Good and Evil in the Færie courts

There are also natural coordination strategies between groups within a mixed simulacrum level blame game. One natural coordination mechanism for a majority (which has some control over which accusations are followed up on) is try to avoid being blamable for anything by only expropriating in "legitimate" ways that have

narrative cover. This allows them to expropriate from others without being punished, and to recruit level-1 players who still take fault analysis literally into their coalition. The price of this coordination strategy is that they can't coordinate overtly. This kind of coalition tends to fly the "good" flag - in Lexical Doll's Seelie and Unseelie Courts paradigm, this is the Seelie Court.

The complement to this coalition is the Unseelie Court, or "Evil," which is willing to be maximally blameworthy. While the Seelie court coordinates to avoid any of its members being blamed, the Unseelie court aestheticizes blameworthiness. Both courts are fundamentally defined by the blame-allocation game.

The "Evil" strategy allows the Unseelie to more overtly coordinate to expropriate from others via mechanisms other than the blame game. Overt coordination - especially on otherwise-unobjectionable things that are simulacrum level 3 crime - makes the Unseelie Court sympathetic to a different class of level 1 players, who see and like that "Evil" is making concrete improvements to the world. The downside of this strategy is that "Evil" is structurally incapable of excluding bad actors, unless it gets big enough that it wants to convert from "Evil" to "Good."

Until recently, Google was "Good" and Uber was "Evil."

"Good" is winning. "Evil" is winning. Who's losing? The level-1 players who just want to fix the things that are wrong and don't want to expropriate from anyone.

****

Related: Talents, Model building and scapegoating

Why exactly is the song 'Baby Shark' so catchy?

(Epistemic status: low-level infohazard, class Earworm. You have been warned.)

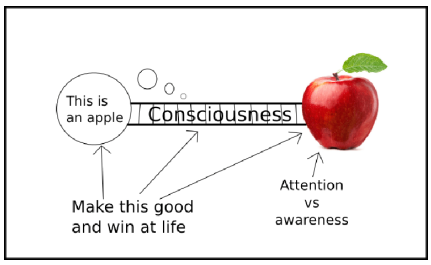
The video, "Baby Shark Dance", has over 2,773,743,743 views as of the time of this post. It bit into me while captive to a parent soothing their child on the BART subway some time ago and has occasionally reared its ugly snout ever since.

I am curious what models of music psychology have to say about this phenomenon and also what those explanations would suggest for increasing the virality of a given piece of music, audiovisuals, or any memetic content in general if applicable.

A quick map of consciousness

Original post: http://bearlamp.com.au/a-quick-map-of-consciousness/

Prior knowledge: Many maps lightly held, Leaky concepts, Boundaries



Map and territory: mind to reality – To be presented alongside the caveat, "what is good?"

(Well "good" is in the map, not the territory. This diagram very quickly becomes a mess, but before that happens, let's talk about reifying the parts of this model to see if it's useful)

To me right now, it seems like consciousness is the ladder between the map and the territory. In the diagram, on the left is a thought, suggesting that "this is an apple" on the right, pictured is a red apple. When the attention points at a red apple, the consciousness is filled with a map of declarative definition that labels, names and concludes that this is an apple.

Consciousness seems to be a label generating machine. Something fundamental about brains is that they map the territory. They quest towards mapping the territory.

That's.Just.What.They.Do.

This brings us to the question of – how do I have a good life. I have 3 strategies:

1. [**content**] Look at different apples

2. [**map**] modify so that there are more positive opinions of apples

3. [**relationality**] appreciate looking at rotten apples if that's what's to look at today.

Content

If I look at dead apples all day, I'm not going to auto-magically have a great day. On the other hand if I look at great apples, I'm going to be impressed and delighted. The apple could be replaced with beautiful artwork, nice sunsets, tasty food, nice music. Whatever strikes in the heart of desire to be attended to. *Improve the content* is a reasonable and helpful strategy sometimes.

Sometimes it’s not the content that’s the problem. Maybe there’s nothing wrong with apples but they make me puke. Then I can try the map.

Map

If every time I see an apple I remember that one time I bit an apple and found half a worm, maybe there’s some work I can do so that I don’t keep thinking worms when I see an apple. Even sunsets are irrelevant when I’m too busy on my phone. If art galleries remind me of my ex, music reminds me of screeching cats (not in a good way), food reminds me of how fat I am (and how I can’t take care of my body). Maybe the work to be done is in the map. Sometimes with more and less force, the map can be trained to be less miserable when presented with stimuli. Usually the good stuff is found by passing through the uncomfortable, not avoiding it.

Sometimes I can’t shift the content. I’m living in the developing world, sometimes sickness and suffering is visible. Sometimes it’s a very real awareness that if I’m not careful it could be me. That’s where the 3rd method comes in.

There’s parts of the map that start to relate to other parts of the map. That’s what I start to call “relationality”.

Relationality

I look at an apple. It reminds me of the time I bit into a worm. How I relate to that content is flexible. I can feel bad about being dumb that time, or I can look at it and laugh about how ridiculous that was. Maybe thinking of worm-apple-gate is my minds way of warning me to be careful it doesn’t happen again. That time I went to see the sunset and could not get off my phone, I was upset about something, maybe I’m being reminded to be kind to myself, now I know better. Screeching cats – Hilarious! Food makes me fat, but it’s really really good food. So tasty! Maybe the question of balancing good food and living!life is worth considering.

I have a chance to see how I’m relating to the content, and I can travel to different maps.

How? Slowly.

That process of “travel to different maps” needs to be done in the way of being that travels all the way down the ladder. If I brute force the attention to move elsewhere, my relationality is “brute force”. My map says, “I gotta brute force my way around here” or “that’s not important” and my content becomes all about the things I avoid. Sure I can brute force my content to be butterflies not machine guns, but that’s not going to substantially change a map with trouble brewing. I can’t always control what I see. but I can work towards relating to those experiences better.

This post has been quick and dirty. I hope to build on it later.

Simple Rules of Law

Response To: Who Likes Simple Rules?

Epistemic Status: Working through examples with varying degrees of confidence, to help us be concrete and eventually generalize.

Robin Hanson has, in his words, “some puzzles” that I will be analyzing. I’ve added letters for reference.

- *A] People are often okay with having either policy A or policy B adopted as the standard policy for all cases. But then they object greatly to a policy of randomly picking A or B in particular cases in order to find out which one works better, and then adopt it for everyone.*
- *B] People don’t like speed and red-light cameras; they prefer human cops who will use discretion. On average people don’t think that speeding enforcement discretion will be used to benefit society, but 3 out of 4 expect that it will benefit them personally. More generally people seem to like a crime law system where at least a dozen different people are authorized to in effect pardon any given person accused of any given crime; most people expect to benefit personally from such discretion.*
- *C] In many European nations citizens send their tax info into the government who then tells them how much tax they owe. But in the US and many other nations, too many people oppose this policy. The most vocal opponents think they benefit personally from being able to pay less than what the government would say they owe.*
- *D] The British National Health Service gets a lot of criticism from choosing treatments by estimating their cost per quality-adjusted-life-year. US folks wouldn’t tolerate such a policy. Critics lobbying to get exceptional treatment say things like “one cannot assume that someone who is wheel-chair bound cannot live as or more happily. … [set] both false limits on healthcare and reducing freedom of choice. … reflects an overly utilitarian approach”*
- *E] There’s long been opposition to using an official value of life parameter in deciding government policies. Juries have also severely punished firms for using such parameters to make firm decisions.*
- *F] In academic departments like mine, we tell new professors that to get tenure they need to publish enough papers in good journals. But we refuse to say how many is enough or which journals count as how good. We’d keep the flexibility to make whatever decision we want at the last minute.*
- *G] People who hire lawyers rarely know their track record and winning vs. losing court cases. The info is public, but so few are interested that it is rarely collected or consulted. People who hire do know the prestige of their schools and employers, and decide based on that.*
- *H] When government leases its land to private parties, sometimes it uses centralized, formal mechanisms, like auctions, and sometimes it uses decentralized and informal mechanisms. People seem to intuitively prefer the latter sort of mechanism, even though the former seems to works better. In one study “auctioned leases generate 67% larger up-front payments … [and were] 44% more productive”.*
- *I] People consistently invest in managed investment funds, which after the management fee consistently return less than index funds, which follow a simple*

clear rule. Investors seem to enjoy bragging about personal connections to people running prestigious investment funds.

- *J] When firms go public via an IPO, they typically pay a bank 7% of their value to manage the process, which is supposedly spent on lobbying others to buy. Google famously used an auction to cut that fee, but banks have succeed in squashing that rebellion. When firms try to sell themselves to other firms to acquire, they typically pay 10% if they are priced at less than $1M, 6-8% if priced $10-30M, and 2-4% if priced over $100M.*
- *K] Most elite colleges decide who to admit via opaque and frequently changing criteria, criteria which allow much discretion by admissions personnel, and criteria about which some communities learn much more than others. Many elites learn to game such systems to give their kids big advantages. While some complain, the system seems stable.*
- *L] In a Twitter poll, the main complaints about my fire-the-CEO decisions markets proposal are that they don't want a simple clear mechanical process to fire CEOs, and they don't want to explicitly say that the firm makes such choices in order to maximize profits. They instead want some people to have discretion on CEO firing, and they want firm goals to be implicit and ambiguous.*

Some of these examples don't require explanation beyond why they are *bad ideas for rules.* They wouldn't work. Others would, or are even obviously correct, and require more explanation. I do think there is a enough of a pattern to be worth trying to explain. People reliably dislike the rule of law, and prefer to substitute the rule of man.

Why do people oppose the rule of law?

You can read his analysis in the original post. His main diagnosis is that people promote discretion for two reasons:

1. They believe they are unusually smart, attractive, charismatic and well-liked, just the sort of people who tend to be favored by informal discretion.
2. They want to signal to others that they have confidence in elites who have discretion, and that they expect to benefit from that discretion.

While I do think these are *related* to some of the key reasons, I do not think these point at the central things going on. Below I tackle all of these cases and their specifics. Overall I think the following are the five key stories:

1. Goodhart's Law. We see a lot of Regressional and Adversarial Goodhart, and also some Extremal and Causal as well. B, F, G, K and L all have large issues here.
2. Copenhagen Interpretation of Ethics. If you put a consideration explicitly into your rule, that makes you blameworthy for interacting with it. And you're providing all the information others need to scapegoat you not only for what you do, but what you *would* do in other scenarios, or *why* you are doing all of this. This is a factor in A, B, D, E, F, H and K all have to worry about this.
3. Forbidden Considerations and Lawsuit Protection. We can take the danger of transparency a step further. There are things that one wants to take into consideration that are illegal to consider, or which you are considered blameworthy for considering. The number of ways one can be sued or boycotted for their decision process goes up every year. You need a complex system in order to disguise hidden factors, and you certainly can't put them into explicit rules. And in general, the less information others have, the less those that don't

like your decision have to justify a lawsuit or making other trouble. B, D, E, F, K and L have this in play.
4. Power. Discretion and complexity favor those with power. If you have power, others worry that anything they do can influence decisions, granting you further power over everything in a vicious cycle. If you use a simple rule, you give up your power, including over unrelated elements. Even if you don't have power directly over a decision, anyone with power anywhere can threaten to use that power whenever any discretion is available, so everyone with power by default favors complex discretionary rules everywhere. This shows up pretty much everywhere other than A, but is especially important in B, F, H, K and L.
5. Theft. Complex rules are often nothing more than a method of expropriation. This is especially central in B, C, D, H, I and J.

I'll now work through the examples. I'll start with C, since it seems out of place, then go in Robin's order.

Long post is long. If a case doesn't interest you, please do skip it, and it's fine to stop here.

C] Tax Returns

The odd policy out here is C, the failure to have the government tell citizens what it already knows about citizens' tax returns. This results in many lost hours searching down records, and much money lost paying for tax preparation. As a commentator points out, the argument that 'this allows one to pay less than they owe' doesn't actually make sense as an explanation. The government still knows what it knows, and still cross-checks that against what you say and pay. In other countries one can still choose to adjust the numbers shared with you by the government.

In theory, one could make a case similar to those I'll make in other places, that telling people what information the government knows and doesn't know allows people to hide anything that the government doesn't know about. But that seems quite minor.

What's going on here is simple regulatory capture, corruption, rent seeking and criminal theft. Robin's link explains this explicitly. Tax preparation corporations like H&R Block are the primary drivers here, because the rules generate more business for them. There is also a secondary problem that fanatical anti-tax conservatives like that taxes annoy people.

But I've never heard of a *regular person* who thinks this policy is a good idea, and I never expect to find one. We're not *this* crazy. We have a dysfunctional government.

A] Random Experiments

Robin's explanations don't fit case A. If you're choosing randomly, no one can benefit from discretion. If you choose the same thing for everyone, again no one can benefit from discretion. If anything, the random system allows participants to potentially cheat or find a way around a selection they dislike, whereas a universal system makes this harder. Other things must be at work here.

This is the opposite of C, a case where people *do* oppose the change, but the change would be obviously good.

I call this the “too important to know” problem.

To me this is a clear case of The Copenhagen Interpretation of Ethics and Asymmetric Justice interacting with sacred values.

An experiment *interacts with the problem* and in particular *interacts with every subject of the experiment, and with every potential intervention,* in a way sufficient to render you blameworthy for not doing more, or not doing the optimal thing.

The contrast between the two cases is clear.

Without an experiment, we’re *forced* to make a choice between options A and B. People *mostly* accept that one must do their best, and potentially sacred values are up against other potentially sacred values, and one must guess and try their best.

In the cases in the study, it’s even more extreme. We’re choosing to implement A or implement B, in a place where normally one would do nothing. So we’re comparing doing something about the situation to doing nothing. It’s no surprise that ‘try to reduce infections’ comes out looking good.

With an experiment, *the choice is between experimentation and non-experimentation.* You are *choosing* to prioritize information *over* all the sacred values the non-objectionable choices are trading off. Even if the two choices are fully non-objectionable, *choosing between them* still means placing *the need to gather information* over *the needs of the people in the experiment.*

The needs of specific people are, everywhere and always, a sacred value. Someone, a particular real person, is on the line. When put up against “information” what chance does this amorphous information have?

Copenhagen explains why it is pointless to say that the experiment is better for these patients than not running one. Asymmetric Justice explains why the benefits to future patients doesn’t make up for it.

There are other reasons, too.

People don’t like their fates coming down to coin flips. They don’t like uncertainty.

People don’t like asymmetry or inequality – if I get A and you get B, *someone* got the better deal, and that’s *not fair*.

If you choose a particular action, that provides evidence that there was a *reason* to choose it. So people instinctively adjust some for the fact that it was chosen. Whereas in an experiment, it’s clear you don’t know which choice is better (unless you *do* know and are simply out to *prove* it, in which case you are a monster). That doesn’t inspire confidence.

A final note is that if you look at the study in question, it suggests another important element. If you choose A, you’re blameworthy for A and for ~B, but you’re certainly not blameworthy for ~A or for B! Whereas if you choose (50% A, 50% B) then you are blameworthy for A, ~A, B *and* ~B, *plus* experimentation in general. That’s a lot of blame.

Remember Asymmetric Justice. If *any* element of what you do is objectionable, *everything* you do, together, is also objectionable. A single ‘problematic’ element

ruins all.

So if we look at Figure 1 in the study, we see in case C that the objection score for the A/B test is actually *below* what we'd expect if we thought the chances of objecting to A and B were independent, and people were objecting to the experiment whenever they disliked either A or B (or both). In cases B and D, we see only a small additional rate of objection. It's only in case A that we see substantial additional objection. Across the data given, it looks like this phenomenon explains about half of the increased rate of objection to the experiments.

It also looks like a lot of people explicitly cited things like 'playing with people's lives' via experiment, and object to experimentation as such at least when the stakes are high.

B] Police Discretion

I do not think Robin's story of expectation of personal benefit is the central story here, either. The correlation isn't even that high in his poll.

> If police have discretion IN GENERAL regarding who they arrest, do you think they will on average use that discretion to arrest those who actually do more net social harm? Do you think that you will tend to be favored by this discretion?
>
> **49%** Yes to both
>
> **13%** No to both
>
> **10%** Yes re net harm, No re me
>
> **28%** No re net harm, Yes re me
>
> — Robin Hanson (@robinhanson) May 6, 2019

If you think net harm is reduced (59%), you're (49/59) 84% to think you'll benefit. If you think net harm is not reduced, you are (28/41) 68% to think you'll benefit. Given that you'd expect models to give correlated returns to the two questions – if discretion is used wisely, it should tend to benefit both most people and a typical civilian looking to avoid doing social harm, and these are Robin's Twitter followers – I don't think personal motivation is explaining much variance here.

The question also asking about only part of the picture. Yes, we would hope that police (and prosecutors and others in the system) would use discretion, at least in part, to arrest those who do more net social harm over those who do less net social harm.

But that's far from the only goal of criminal justice, or of punishment. We would *also* hope that authorities would use discretion to accomplish other goals, as well.

Some of these goals are good. Others, not so much.

What are some other goals we might have? What are other reasons to use discretion?

I think there are five broad reasons, beyond using it to judge social harm.

1. Discretion gives law enforcement authorities power. This power allows them to maintain and exert authority. It keeps the system running, ensures cooperation and allows them to solve cases and enforce the law.
2. Discretion gives law enforcement authorities power and status. This power and status is a large part of the compensation we give to law enforcement.
3. Discretion allows those with status and/or power to gain additional benefits from that status and/or power.
4. Discretion allows us to enforce rules without having to state those rules explicitly, be able to well-specify those rules, or specify them in advance.
5. Discretion guards against Goodhart's Law and the gaming of the system.

To this we would add Robin's explanations, that one might want to benefit from this directly, and/or one might want to signal support for such authorities. And the more discretion they have, the more one would want to signal one's support – see reason 1.

A case worth grappling with can be made *for or against* each of these five justifications being net good. So one could argue in favor like this with arguments like (I am not endorsing these, nor do they necessarily argue for today's American level of discretion):

1. No one would cooperate with authorities. Every potential witness will stonewall, every defendant will fight to the end. Intimidation of witnesses would be impossible to stop. Good luck getting anyone even *associated with someone* doing *anything* shady to ever talk to the police. Cases wouldn't get solved, enforcement would break down. Criminals would use Blackmail to take control of people and put them outside the law. Authorities' hands would be tied and we'd be living in Batman's Gotham.
2. In many ways, being a cop is a pretty terrible job. Who wants to constantly have hostile and potentially dangerous interactions with criminals? Others having to 'respect my authoritah' improves those interactions, and goes a long way in making up for things on many fronts.
3. Increasing returns for status and power increase the power of incentives. We want people to do things we approve of, and strive to gain society's approval, so we need levers to reward those who do, and punish those who don't. We want to reward those who walk the straight and narrow track, and make good, and also those who strive and achieve. This has to balance the lure of anti-social or criminal activity, especially in poorer communities. And it has to balance the lure of inactivity, as we provide increasingly livable baskets of goods to those who opt out of status and prestige fights to play video games.
4. If we state our rules explicitly, we are now blameworthy for every possible corner case in which those rules result in something perverse, and for every motivation or factor we state. Given that we can be condemned for each of these via Asymmetric Justice, this isn't an option. We also need to be able to implement systems without being able to specify them and their implications in an unambiguous way. If we did go by the letter of the law, then we're at the mercy of the mistakes of the legislature, and the law would get even more complex than it is as we attempt to codify every case, beyond the reach of any person to know how it works. Humans are much better at processing other types of complex systems. And we need ways for local areas to enforce norms and do locally appropriate things, without having their own legislatures. If the system couldn't correct obvious errors, both Type I and Type II, then it would lose the public trust. And so on.
5. If people know exactly what is and isn't illegal in every case, and what the punishments are, then it will be open season for anything you didn't make

explicitly illegal. People will find technically permitted ways to do all sorts of bad things. They'll optimize for all the wrong things. Even when they don't, they'll do the maximum amount of every bad thing that your system doesn't punish enough to stop them from doing, and justify it by saying 'it was legal'. Your laws won't accomplish what they set out to accomplish. Any complex system dealing with the messiness of the real world needs some way to enforce the 'spirit of the law.' One would be shocked if extensive gaming of the system and epic Goodhart failures didn't result from a lack of discretion.

Or, to take the opposite stances:

1. Power corrupts. It is bad and should be minimized. The system will use its power to protect and gather more power for itself, and rule for the benefit of its members and those pulling their strings, not you. If we didn't have discretion, we could minimize the number of laws and everyone wouldn't need to go around afraid of authority. And people would see that authority was legitimate, and would be inclined to cooperate.
2. Power not only corrupts, it also attracts the corrupt. If being a cop lets you help people and stand for what's right, and gives you a good and steady paycheck, you'll get good people. If being a cop is an otherwise crappy job where you get to force people to 'respect my authoritah' and make bank with corrupt deals and spurious overtime, you'll get exactly the people you don't want, with exactly the authority you don't want them to have.
3. The last thing we want is to take the powerful (and hence corrupt) and give them even more discretion and power, and less accountability. The inevitable result is increasing amounts of scapegoating and expropriation. This is how the little guy, and the one minding their own business trying to produce, get crushed by corporations and petty tyrants of all sorts.
4. If you can't state what your rules are, how am I supposed to follow them? How can it be fair to punish me for violating rules you won't tell me? This isn't rules at all, or rule of law. It's rule of man, rule of power begetting power. The system is increasingly rigged, and you're destroying the records to prevent anyone from knowing that you're rigging it.
5. The things you're trying to protect from being gamed are power and the powerful, which are what are going to determine what the 'spirit of the rules' is going to be. So the spirit becomes whatever is good for them. The point of rule of law, of particular well-specified rules, is to avoid this. Those rules are not supposed to be the only thing one cares about. Rather, the law is supposed to guard against specific bad things, and other mechanisms allowed to take care of the rest. If the law is resulting in Goodhart issues, it's a bad law and you should change it.

The best arguments I know about against discretion have nothing to do with the social harm caused by punished actions. They are arguments for rule of law, and to guard against what those with discretion will do with that power. These effects are rather important and problematic even when the system is working as designed.

The best arguments I know about in favor of discretion also have nothing to do with the social harm caused by punished actions. They have to do with the system depending on discretion in order to be able to function, and in order to ensure cooperation. A system without discretion by default makes the spread of any local information everyone's enemy, and provides no leverage to overcome that. If we didn't have discretion, we would have to radically re-examine all of our laws and our entire system of enforcement, lest everything fall apart.

My model says that we currently give authorities too much discretion, and (partly) as a result have punishments that are too harsh. And also that the authorities have so much discretion partly because punishments have been made too harsh. Since discretion and large punishments give those with power more power, it would be surprising if this were not the case.

D] National Health Service

The National Health Service gets criticized constantly because it is their job to deny people health care. There is not enough money to provide what we would think of as an acceptable level of care under all circumstances, because our concept of acceptable level of care is *all of the health care.* In such a circumstance, there isn't much they could do.

Using deterministic rules based on numbers is the obviously correct way to ration care. Using human discretion in each case will mean either always giving out care, since the choice is between care or no care – which is a lot of why health care costs are so high and going higher – or not always giving out care when able to do so, which will have people screaming unusually literal bloody murder.

Deterministic rules let individuals avoid blame, and allow health care budgets to be used *at all.* But that doesn't mean people are going to like it. If anything, they're going to be mad about both the rules and the fact that they don't have a human they can either blame or try to leverage to get what they want. There's also the issue of putting a value on human life at all, which is bad enough but clearly unavoidable.

More than that, once you explicitly say what you value by putting numbers on lives and improvements in quality of life, you're doing something both completely necessary and completely unacceptable. The example of someone in a wheelchair is pretty great. If you don't provide *some* discount in value of quality of life for physical disability, then you are saying that physical disabilities don't decrease quality of life. Which has pretty terrible implications for a health care system trying to prevent physical disabilities. If you do say they decrease quality of life, you're saying people with disabilities have less value. There are tons of places like this.

Another way to view this is that the only way for one to make health care decisions to ration care or otherwise sacrifice sacred values to stay on budget, without blame, is to have all those decisions be seen as out of your control and not your choice. The only known way to do that is to have a system in place, and point to that. That system then becomes a way to *not* interact with the system, avoiding blame. Whereas proposing or considering any *other* system involves interaction, and thus blame.

E] Value of Life

If you are caught making a trade-off between a sacred value (life) and a non-sacred value (money), it's not going to go well. Of course a company doing an explicit calculation here is going to get punished, as is a government policy making an explicit comparison. Humans don't care about the transitive property.

Thus, firms and governments, who *obviously* need to value risk to human life at a high but finite numerical cost, will need to do this without writing the number down explicitly in any way. This is one of the more silly things one cannot consider, that one obviously *must* consider. In a world where we are blameworthy (to the point of being

sued for massive amounts) for doing explicit calculations that acknowledge trade-offs or important facts about the world, firms and governments are forced to make their decisions in increasingly opaque ways. One of those opaque preferences will be to favor those who rely on opaqueness and destroy records, and to get rid of anyone who is transparent about their thinking or otherwise, and keeps accurate records.

F] Tenure Requirements

Tenure is about evaluating what a potential professor would bring to the university. No matter what extent politics gets involved, this is someone you'll have to work with for decades. After this, rule of law *does* attach. You won't be able to fire them afterwards unless they violate one of a few well-defined rules – or at least, that's how it's *supposed* to work, to protect academic freedom, whether or not it does work that way. You'll be counting on them to choose and do research, pursue funding, teach and advise, and help run the school, and be playing politics with them.

That's a big commitment. There are lots more people who want it and are qualified on paper than there are slots we can fund. And there are a lot more things that matter than how much research one can do. Some of them are things that are illegal to consider, or would look bad if you were found to be considering them. Others simply are not research done. You can't use a formula, because people bring unique strengths and weaknesses, and you're facing other employers who consider these factors. Even if a simple system could afford to mostly 'take its licks' you would face massive adverse selection, as everyone with bad intangibles would knock at your door.

You need to hold power over the new employees, so they'll do the work that tenured employees don't want to do, and so they'll care about all aspects of their job, rather than doing the bare technical minimum everywhere but research.

Then there's the Goodhart factors on the papers directly. One must consider how the publications themselves would be gamed. If there was a threshold requirement for journal quality, the easiest journals that count would be the only place anything would be published. If you have a point system, they'd game that system, and spend considerable time doing it. If you don't evaluate paper quality or value, they won't care at all about those factors, focusing purely on being good enough to make it into a qualifying journal. Plus, being able to evaluate these questions *yourself* without an outside guide or authority will be part of the job you're trying to get. We need to test that, too.

What you're *really* testing for when you consider tenure, ideally, is not only skill but also *virtue.* You want someone who is *naturally driven* to scholarship and the academy, to drive forward towards important things. While also caring enough to do a passable job with other factors. Otherwise, once they can't be fired, you won't be able to get them to do anything. Testing for virtue isn't something you can quantify. You want someone who will aim for the spirit rather than the letter, and who knows what the spirit is and cares about it intrinsically. If you judge by the letter, you'll select for the opposite, and if you specify that explicitly, you'll lose your signal that way as well.

I'd write this one up to power and exploitation of those lower on the totem pole, the need to test for factors that you can't say out loud, the need to test for virtue, and the need to test for knowing what is valuable.

G] A Good Lawyer

People rightfully don't think this number will tell us much, even now when it is not being gamed and vulnerable to Goodhart. Robin seems to be assuming that one should think that a previous win percentage should be predictive of a lawyer's ability to win a particular case, rather than being primarily a selection effect, or a function of when they settle cases.

I doubt this is the case, *even with a relatively low level of adversarial Goodhart effects.*

Most lawyers or at least their firms have great flexibility in what cases they pursue and accept. They also have broad flexibility in how and when they settle those cases, as clients largely rely on lawyers to tell them when to settle. Some of them will mostly want cases that are easy wins, and settle cases that likely lose. Others, probably better lawyers for winning difficult cases, will take on more difficult cases and be willing to roll the dice rather than settle them.

I don't even know what counts as a 'win' in a legal proceeding. In a civil case you strategically choose what to ask for, which might have little relation to realistic expectations for a verdict, so getting a lesser amount might or might not be a 'win' and any settlement might be a win or loss even if you know the terms, and often the terms are confidential.

Thus, if I was looking for a lawyer, I would continue to rely on personal recommendations, especially from lawyers I trust, rather than look at overall track records, even if those track records were easily available. I don't think those track records are predictive. Asking questions like someone's success in similar style cases, with richer detail in each case, seems better, but one has to pay careful attention.

If people started using *win-loss records* to *choose lawyers,* and lawyers started optimizing their win-loss records, what little information those records might have gets even less useful. You would mostly be measuring which lawyers prioritize win-loss records, by selecting winners and forcing them to verdict, while avoiding, settling or pawning off losers, and by getting onto winning teams, and so on. By manipulating the client and getting them to do what was necessary. It's not like lawyers don't mostly know which cases are winners. By choosing a lawyer with too good a win-loss record, you'd be getting someone who cares more about how they look in a statistic than doing what's right for their clients, and also who has the flexibility to choose which cases they have to take.

The adverse selection here, it burns.

That's what I'd actually expect now. Some lawyers *do* care a lot about their track records, they'll have better track records, and they're exactly who you want to avoid. I'd take anyone bragging about their win rate as a very negative sign, not a positive one.

So I don't think this is about simple rules, or about people's cognitive errors, or anything like that. I think Robin is just proposing a terrible measure that is not accurate, not well-defined and easily gamed, and asking why we aren't making it available and using it.

Contrast this with evaluations of doctors or hospitals for success rates or death rates from particular surgeries. That strikes me as a *much better* place to implement such strategies, although they still have big problems with adversarial Goodhart if you started looking. But you can get a much better idea of what challenges are being tackled and about how hard they are, and a much better measure of the rate of success. I'd still worry a lot about doctors selecting easy cases and avoiding hard ones, both for manipulation and because of what it would do to patient care.

A general theme of simple rules is that when you reward and punish based on simple rules, one of the things you are rewarding is a willingness to prioritize maximizing for the simple rule over any other goal, including the thing you're trying to measure. Just like any other rule you might use to reward and punish. The problem with simple rules is that they explicitly shut out one's ability to notice such optimization and punish it, which is the natural way to keep such actions in check. Without it, you risk driving out anyone who cares about anything but themselves and gaming the system, and creating a culture where gaming the system and caring about yourself are the only virtues.

H] Land Allocations

If all you care about is the 'productivity' of the asset and/or the revenue raised, then *of course* you use an auction. Easy enough, and I think people recognize this. They *don't want* that. They want a previously public asset to be used in ways the public prefers, and think that we should prefer some uses to other uses because of the externalities they create.

It seems reasonable to use the opportunity of selling previously public goods to advance public policy goals that would otherwise require confiscating private property. Private sellers will also often attach requirements to sales, or choose one buyer over another, sacrificing productivity and revenue for other factors they care about.

We can point out all we like how markets create more production and more revenue, but we can't tell people that they should care mostly about the quantity of production and revenue instead of other things. When there are assets with large public policy implications and externalities to consider, like the spectrum, it makes sense to think about monopoly and oligopoly issues, about what use the assets will be put to by various buyers, and what we want the world to look like.

That doesn't mean that these good factors are the primary justifications. If they were, you'd see conditional contracts and the like more often, rather than private deals. The real reason is *usually* that other mechanisms allow insiders to extract public resources for private gains. This is largely a story of brazen corruption and theft. But if we're going to argue for simple rules because they maximize simple priorities, we need to also argue for why those priorities cover what we care about, or we'll be seen as tone deaf at best, allowing the corrupt to win the argument and steal our money.

I] Investment Funds

Low fee index funds are growing increasingly popular each year, taking in more money and a greater share of assets. Their market share is so large that being included in a relevant index has a meaningful impact on share prices.

Managed funds are on the decline. Most of these funds are *not* especially prestigious and most people invested in them don't brag about them, nor do they have much special faith in those running the funds. They're just not enough on the ball to realize they're being taken for a ride by professional thieves.

Nor do I think most people care about associating with high status hedge funds or anything like that. I don't see it, at all.

Also, those simple rules? You can find them in active funds, too. A lot of them are pretty popular. Simple technical analysis, simple momentum, simple value rules, and so on. What counts as simple? That's a matter of perspective. Index providers are often doing staggeringly complex things under the hood. And indexing off someone else's work is a magician's trick, free riding off the work of others in a way that gets dangerous if too many start relying on it.

Most regular investors who think about what they're doing at all, know they should likely be in index-style funds, and increasingly that's where they are. If there's a mystery at all it's at least contained at the high end, in hedge funds with large minimums.

One can split the remaining 'mystery' into two halves. One is, why do some people think there exist funds that have sufficient alpha to justify their fees? Two is, why do some people think they've found one of those funds?

The first mystery is simple. They're right. There exist funds that have alpha, and *predictably* beat the market. The trick is finding them and getting your money in (or the even better trick is figuring out how to do it yourself). I don't want to get into an argument over efficient markets here and won't discuss it in the comments, but the world in which no one can beat the market *doesn't actually make any sense.*

The second mystery is also simple. Marketing, winners curse, fooled by randomness and adverse selection, and the laws of markets. Of course a lot more people think they've found the winner than have actually found one.

This is a weird case in many ways, but my core take here is that the part of this that *does* belong on this list, is an example of complexity as justification for theft.

J] Selling the Company

Google is the auction company. They were uniquely qualified to run an auction and bypass the banks, and did it (as I understand it) largely because it was on brand and they'd have felt terrible doing otherwise. A more interesting case is Spotify, who recently simply let people start trading its stock without an IPO at all. Although they still paid the banks fees, which I find super weird and don't understand. There never was a rebellion.

How do banks extract the money?

My model is something like this, coming mostly from reading Matt Levine. The banks claim that they provide essential services. They find and line up customers to buy the stock, they vouch for the stock, they price the stock properly to ensure a nice bump so everyone feels happy, they backstop things in case something goes wrong, they handle a ton of details.

What they really do are two things. Both are centered around the general spreading by banks of FUD: Fear, Uncertainty and Doubt.

First, they prevent firms from suddenly having to navigate a legally tricky and potentially risky, and potentially quite complex, world they know nothing about, where messing up could be a disaster. One *does not simply* sell the company or take it public, as much as it might look simple from the outside. And while the bank's marginal costs are way, way lower than what they charge, trying to get that expertise in house in a confident way is hard.

Second, they are what people are *comfortable* with. You're *not blameworthy* for paying the bank. It's the null action. If you do it, no one says 'hey they've robbed us all of a huge amount of money.' Instead, they say 'good on you for not being too greedy and trying to maximize the price while risking the company's future.'

They're doing this at the crucial moment when *how you look* is of crucial importance, when you're about to get a huge windfall for years or a lifetime of work and give the same to everyone who helped you. When you're spending all your energy negotiating lots of other stuff. A disruption threatens to unravel all of that. What's a few percent in that situation? So what if you don't price your IPO as high as you could have so that bankers can enjoy their bounce?

Banks are conspiring with the buyers to cheat the sellers out of the value of what they bring to the table. Buyers who object are threatened with ostracism and being someone no one is *comfortable* with, with the other side walking away from the table after buyers put in the work to get here.

Is this all guillotine-worthy highway robbery? Hell yes. Completely.

Banks (and the buyers who are their best customers and allies) are colluding with this pricing, and that's the *nicest* way to put this. Again, this is *theft.* Complexity is introduced to allow rent seeking and theft, exploiting a moment of vulnerability.

K] College Admissions

Interesting that Robin says the system 'appears stable.' To me it does *not* seem stable. We just had a *huge* college admissions scandal that damaged faith in the system and a quite-well justified lawsuit against Harvard. We have the SAT promising to introduce 'adversity scores.' We have increasingly selective admissions eating more and more of childhood, and the rule that what can't go on forever, won't. This calls for some popcorn.

What's causing the system to be complex? We see several of the answers in play here.

We see the 'factors you can't cite explicitly' problem and the 'we don't want something we can be sued or blamed for' here. Admissions officers are trying to pick kids who will be successful at school and in life, as well as satisfy other goals. A lot of the things that predict good outcomes in life are not things you would want to be caught dead using as a determinant in admissions even if they weren't illegal to use in admissions. The only solution is to make the system complex and opaque, so no one can prove what you were thinking.

We also see complexity as a way for the rich and powerful to expropriate resources, in the sense that the rich and powerful and their children are likely to be more successful, and more likely to give money to the school. And of course, if the school has discretion, that gives the school power. It can extract resources and prestige from others who want to get their kids in. Employees, especially high-up ones, can extract things even without illegal bribes. Why pass that up?

We see the Goodhart's Law and adverse selection problems. If you admit purely on the basis of a test, and the other schools admit on the basis of a range of factors, you don't get the best test scorers unless you're Harvard. You get the kids who are an epic fail at those other factors.

If you give kids an explicit target, they and their parents will *structure their entire lives* around it. They'll do that even with a vague implicit target, *as they do now.* If it's explicit, you get things like you see in China, where (according to an eyewitness who once came to dinner) many kids are pulled from school and do nothing but cram facts into their heads for the college admissions exam for *years.* And why shouldn't they?

So you get kids whose real educations are crippled, who have no life experience and no joy of childhood. The only alternative is to allow a general sense of who the kid is and what they've done to matter. To be able to holistically judge kids and properly adjust.

As always, the more complex and hard to understand the game, the greater the expert's advantage. The rich and powerful who understand the system and can make themselves look good will have a large edge from that. And the more we explicitly penalize them for those advantages, but not for their gaming of the system, the more we *force* them to game the system even harder. If you use an adversity score to set impossibly high standards for rich kids, they're going to use every advantage they have to make up for that even more than they already do.

And of course, part of the test is seeing how you learn to game the test and what approach you take. Can you do it with grace? Do you do too much of it, not enough or the right amount?

This is all an anti-inductive arms race. The art of gaming the system is in large part the art of making it look like you're not gaming the system, which is an argument for simpler rules. At this point, what portion of successful admissions is twisting the truth? How much is flat out lying? How much is presenting yourself in a misleading light? To what extent are we training kids from an early age to have high simulacrum levels and sacrifice their integrity? A lot. Integrity being explicitly on the test just makes it one more thing you need to learn to fake.

I hate the current situation, and the educational system in general, but I think the alternative of a simple, single written test, with the system otherwise unchanged, would be *worse.* But of course, we'd never let it be that simple. That's all before the fights over how to adjust those scores for 'adversity' and 'diversity,' and how to quantify that, and the other things we'd want to factor in. Can you imagine what happens to high schools if grades don't matter? What if grades did matter in a *formulaic* matter and students and teachers were forced to confront the incentives? The endless battles over what other life activities should score points, the death of any that don't, and the twisting into point machines of those that do?

So here we have all the Goodhart problems, and the theft problems, and the power problems, and the blameworthy considerations and justifications problems and lawsuit

problems with their incentive to destroy all information. The gang's all here.

L] Fire that CEO

I love me a prediction market, but you have to do it right. Would enough people and money participate? If they did, would they have the right incentives? If both of those, would you want that to be how you make decisions?

I think the answer to the first question is yes, if you structure it right. If there are only two possibilities and one of them will happen, you can make it work.

The answer to the second question is, no.

We can consider two possibilities.

In scenario one, this acts as an *advisory* for the board, to help them decide what to do.

In scenario two, this is the sole thing looked at, and CEOs are fired if and only if they are judged to be bad for the stock price, or can otherwise only be fired for specific causes (e.g. if found shooting a man on Fifth Avenue or stealing millions in company funds and spending them on free to play games, you need to pull the plug without stopping to look at the market).

The problem with scenario one is that you're trading on how much the company is worth *given that the CEO was fired.* That's *very different* from what you think the company would be worth if we decided to fire the CEO. The scenarios where the CEO is fired are where the board is unhappy with them, which is usually because of bad things that would make us think the stock is likely to be less valuable, like the stock price having gone down or insiders knowing the CEO has done or will do hard to measure long term damage. That doesn't mean it won't *also* take into account other things like whether the CEO is paying off the board, but the correlation we're worried about is still super high. Giving the board discretion, that market participants would expect the board to use, hopelessly muddles things.

You could try to solve that problem by having the market trade only very close in time to the board decision. You kind of have to do that anyway, to avoid having a lame duck CEO. But it still depends on a lot of private information, and the decision will still reveal a lot about the firm. So I think that realistically this won't work.

The problem with scenario two is that you've taken away any ability to punish or reward the CEO for anything other than future stock prices. This effectively gives the CEO absolute power, and allows them to get away with any and all bad behavior of all kinds. Even if past behavior lowers the stock price, it only matters to the extent that it predicts future actions which would further lower the stock price. So CEOs don't even have to care about the stock price. They only need to care about the stock price predictions in relation to each other. So the best thing the CEO can do is make getting rid of them as painful as possible. Even more than now, they want to make sure that losing them destroys the company as much as possible. Their primary jobs now are to hype themselves as much as possible to outsiders, and to spend capital manipulating these prediction markets.

Again, we're seeing Goodhart problems, we're seeing reinforcement of power (in this case, of the board over the CEO, so it's a balance of power we likely welcome), and the ability to take things into consideration without needing to make them explicit or

measurable, as companies both care about things they’re not legally allowed to care about and which we wouldn’t like hearing they cared about, especially explicitly, and they need to maintain confidentiality.

By default, avoid ambiguous distant situations

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

> "*The Ood.[...] They're born for it. Basic slave race.*" Mr. Jefferson, *The Impossible Planet*.
>
> "*And although he may be poor, he shall never be a slave,*", from the *Battle Cry of Freedom*.

I've talked about morally distant situations: situations far removed from our own. This distance is in the moral sense, not the actual sense: ancient China is further from us than Star Wars is (even for modern Chinese). There are possible worlds out there far more alien than anything in most of our fiction.

In these distant situations, our usual web of connotations falls apart, and closely related terms start to mean different things. As shown in this post, in extreme cases, our preferences can become nonsensical.

Note that not all our preferences need become nonsensical, just some of them. Consider the case of a slave race - a willing slave race. In that situation, our connotations about slavery may come apart: willing and slave don't fit well together. However, we would still be clear ourselves that *we* didn't want to be slaves. So though some preferences lose power, some do not.

But let's return to that willing slave race situation. Eliezer's Harry says:

> Whoever had created house elves in the first place had been unspeakably evil, obviously; but that didn't mean Hermione was doing the right thing now by denying sentient beings the drudgery they had been shaped to enjoy.)

Or consider Douglas Adam cow-variant bread for willingly being eaten:

> "I just don't want to eat an animal that's standing here inviting me to," said Arthur, "it's heartless."
>
> "Better than eating an animal that doesn't want to be eaten," said Zaphod. [...]
>
> "May I urge you to consider my liver?" asked the animal, "it must be very rich and tender by now, I've been force-feeding myself for months."

At X, the decision is clear. Should we go to X in the first place?

In those situations, some immediate actions are pretty clear. There is no point in freeing a willing slave race; there is no advantage to eating an animal that doesn't want to be eaten, rather than one that does.

The longer-term actions are more ambiguous, especially as they conflict with other of our values: for example, should we forcibly change the preferences of the slave race/edible race so that they don't have those odd preferences any more? Does it make a difference if there are more manipulative paths that achieve the same results, without directly forcing them? We may not want to allow manipulative paths to count as acceptable in general.

But, laying that aside, it seems there is a prima facie case that we shouldn't enter those kinds of situations. That non-conscious robots are better than conscious willing slaves. That vat grown meat is better than conscious willing livestock.

So there seems to be a good rule of thumb: don't go there. Add an axiom A:

- A: When the web of connotations of a strong preference falls apart, those are situations which should get an automatic penalty. Initially at least, those should be treated as bad situations worth avoiding.

Default weights in distant situations

When a web of connotation unravels, the preferences normally end up weaker than initially, because some of the connotations of those preferences are lost or even opposite. So, normally, preferences in these distant situations are quite weak.

But here I'm suggesting adding an explicit meta-preference to these situations. And one that the human subject might not have themselves. This doesn't fit in the formalism of this post. In the language of the forthcoming research agenda, this is a "Global meta-preferences about the outcome of the synthesis process".

Isn't this an overriding of the person's preference? It is, to some extent. But note the "Initially at least" clause in A. If we don't have other preferences about the distant situation, it should be avoided. But this penalty can be overcome by other considerations.

For example, the previous standard web of connotations for sexuality has fallen apart, while the gender one is unravelling; it's perfectly possible to have meta-preferences that would have told us to respect our reflection on issues like that, and our reflection might be fine with these new situations. Similarly, some (but not all) future changes to the human condition are things that would worry me initially but that I'd be ok with upon reflection; I myself have strong meta-preferences that these should be acceptable.

But for situations where our other preferences and meta-preferences don't weigh in, A would downgrade these distant worlds as a default (dis)preference. This adds an explicit level of status quo bias to our preferences, which I feel is justified: better to be prudent rather than reckless where our preferences and values are concerned. The time for (potential) recklessness is in the implementation of these values, not their definition.

How much do major foundations grant per hour of staff time?

[Edit: original source has been found. The way I framed this question was off, since I'd mis-remembered it as "amount that a given grantmaker grants per year" rather than "grants per hour of staff time". Updated the title]

I recall reading an article once that claimed that, when examining many small and large foundations, it turned out that there was a maximum amount that a given grantmaker typically gave out. And as organizations scaled to give out more money, this amount stayed surprisingly fixed, with a higher overhead ratio than you might have expected.

(i.e. when an org gives out a million a year, it has N grantmakers, and when it gives out 100 million a year, it typically has 100N grantmakers).

I don't remember the number, or the methodology that determined it. Curious if anyone can remember the article. (It might have been from OpenPhil's blog, or it might have been some random news site).

I vaguely remember the number "3" being involved, possibly $300k, or $3 million.

The takeaway I remember was something like "you might naively think you can scale up an organization and then give away money more efficiently, but weird forces seem to limit that."

Does this sound familiar to anyone?

April 2019 gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2019/04

Schelling Fences versus Marginal Thinking

Follow-up / Related to: Scott Alexander's Schelling Fences on Slippery Slopes, Sunk Cost Fallacy, Gwern's Are Sunk Costs Fallacies?, and Unenumerated's Proxy Measures, Sunk Costs, and Chesterton's Fence

I was recently reading an essay by Clayton Christensen, in the (fairly worthwhile) HBR's "Must Reads" boxed set, where he recommends that people "Avoid the Marginal Cost Mistake". In short, he suggests that Schelling Fences are sometimes ignored, or not constructed, because of a somewhat fallacious application of marginal-cost thinking. For example, my Schelling fence for work is that I stop when it is time to get my kids. The other side is that occasionally I'm in the middle of something - coding, or writing this lesswrong post - where being interrupted is fairly high cost. I can usually ask someone else to pick them up instead, and given how much I see them, the marginal value of time with my kids is low.

Christensen suggests that this analysis is incorrect, largely because of myopia. I am ignoring the longer term benefits of family dinners because the connection between coming home today and building the norm of being home for dinner every night is a longer-term investment. The future is full of extenuating circumstances, and only a fairly strong Schelling fence will let me insist that my kids stay home for dinner once they are teenagers.

I'd apply it more broadly, but his point was that this is especially critical in matters of morality. Cheating once changes everything. The simple fact that you cheated weakens your resolve not to in the future. The spiral created by a single action leads easily down a path towards using infinite money and invulnerability cheat codes, with no further challenge or enjoyment from playing the video game - or in the context he's discussing, it led to jail time for two of the people from his graduating class back in college.

Conclusions?

The critical question is: where do we want to use marginal cost analysis, and where do we want to stick to our sunk-costs and Schelling fences?

Based on Christensen's analysis, I would suggest that Schelling fences rather than sunk costs are particularly valuable for reinforcing values that are hard to measure, are too long term to get routine feedback on, or that involve specific commitments to other people. On the other hand, based on Gwern's work, I think there are places where marginal costs are under-appreciated, especially in relation to other people. Below, I lay out some settings and examples on each side.

Some examples of where to consider reinforcing fences and avoiding simplistic marginal cost thinking might include:

- Going to a weekly meet-up that reinforces your connections to a good epistemic community and/or effective altruist values. Value drift is a long-term concern that needs short term reinforcement.
- Anything involving family or long-term relationships. Marginal cost thinking is poisonous for relationships, since the benefits of investing in the relationship are not very visible, and long term.
- Moral rules. Utilitarian and consequentialist thinking is easy to use to make yourself stupider. At the very least, you should be asking others - just like this is useful to avoid unilateralist curses, it is useful to avoid self-deception and convenient excuses.
- Where there are switching costs or longer term goals. Learning to play guitar instead of continuing to practice piano (or moving from C++ to Python) is easy to justify in the short term, but expensive in terms of changes needed and resetting progress.
- When goals are unknown. As Unenumerated put it, "cases where substantial evidence or shared preferences that motivated the original investment decision have been forgotten or have not been communicated, or otherwise where the quality of evidence that led to that decision may outweigh the quality of evidence that is motivating one to change one's mind."

Some examples of where it seems useful to avoid constructing Schelling fences, and to try paying more attention to marginal cost:

- When constructing rules for other people, or in orgnaizations. Schelling fences are useful for self-commitment, otherwise they are rules and formal structures rather than norm-based fences. As gwern noted, " Whatever pressures and feedback loops cause sunk cost fallacy in organizations may be *completely* different from the causes in individuals."
- When the environment is very volatile, and non-terminal goals change. It's easy to get stuck in a mode where the justification is "this is what I do," rather than a re-commitment to the longer term goal. If you are unsure, try revisiting why the fence was put there. (But if you don't know, be careful of removing Chesterton's Fence! See "When goals are unknown", above.)
- When the fence is based on a measurable output, rather than an input. In such a case, the goal has been reified, and is subject to Goodhart effects. Schelling fences are not appropriate for outcomes, since the outcome isn't controlled directly. (Bounds on outcomes also implicitly discourage further investment - see: Shorrock's Law of Limits. If necessary, the outcome itself should be rewarded, rather than fenced in.)

Episode 5 of Tsuyoku Naritai (the 'becoming stronger' podcast): The Stance

Transcript in show notes, as always!

https://anchor.fm/tsuyokunaritai/episodes/Episode-5---The-Stance-e434pj

Does the Higgs-boson exist?

This is a linkpost for https://backreaction.blogspot.com/2019/05/does-higgs-boson-exist.html

> What do scientists mean when they say that something exists? Every time I give a public lecture, someone will come and inform me that black holes don't exist, or quarks don't exist, or time doesn't exist. Last time someone asked me "Do you really believe that gravitational waves exist?"

Sabine is a theoretical physicists who had gained prominence (and notoriety) through her book Lost in Math, about groupthink in high-energy physics.

In this post she sums up beautifully what I and many physicists believe, and is vehemently opposed by the prevailing realist crowd here on LW. A few excerpts:

> Look, I am a scientist. Scientists don't deal with beliefs. They deal with data and hypotheses. Science is about knowledge and facts, not about beliefs.

...

> We use this mathematics to make predictions. The predictions agree with measurements. That is what we mean when we say "quarks exist": We mean that the predictions obtained with the hypothesis agrees with observations.

...

> Now, you may complain that this is not what you mean by "existence". You may insist that you want to know whether it is "real" or "true". I do not know what it means for something to be "real" or "true." You will have to consult a philosopher on that. They will offer you a variety of options, that you may or may not find plausible.
>
> A lot of scientists, for example, subscribe knowingly or unknowingly to a philosophy called "realism" which means that they believe a successful theory is not merely a tool to obtain predictions, but that its elements have an additional property that you can call "true" or "real". I am loosely speaking here, because there several variants of realism. But they have in common that the elements of the theory are more than just tools.
>
> And this is all well and fine, but realism is a philosophy. It's a belief system, and science does not tell you whether it is correct.

...

> Here is a homework assignment: Do you think that I exist? And what do you even mean by that?

Programming Languages For AI

Firstly a chess analogy

Suppose that you were part of a team trying to build a chess playing program. Your team has not yet had the fundamental insight of min-max search with approximate evaluations. While you definitely want some people on the team thinking hard about the abstract nature of chess, this is a somewhat serial process, is there anything else that could usefully be done in parallel?

While we wouldn't have the insights to no exactly what a chess engine would look like, we can say some things about the code. The code will almost certainly want some sort of representation of chess pieces or chess boards. So implementing a virtual chessboard would not be a waste of time.

More speculatively, you might overhear the **Thinking Hard About Chess** department talk about how a good chess position was defined in terms of making the opponents move not good, and invent recursion.

Ideally, when the crucial insights have been had, all the building blocks needed to make it are ready to go. The first chess engine is half a page of *chesslang* code.

On to AI

So, can we think of any programmatic building blocks that are likely to be useful in building an AI. Yes, arithmatic, if statements, lists and so on.

These features have already been implemented in many programming languages, can we think of any features that might be useful in making AI that aren't easily available in any programming languages?

Much current AI is arithmetical with neural networks and propagation. However, there are already several fairly easy to use libraries for this, and I can't see how to make it substantially easier to program this sort of AI, other than stuff like hyperparameter optimization that lots of people are already working on.

Much theoretical research on AI is symbolic, even Godelian. AIXItl for example, executes every piece of code up to a certain length and runtime. That would suggest that the programming language should make it easy to generate syntactically correct code, and to run it for a bounded time without side effects. This is also what you would want if you were generating code with an evolutionary algorithm.

On the alignment forum, several designs of AI mentioned involve "search for a proof that this piece of code halts". So our programming language for AI should have powerful proof handling facilities.

Suggested Mechanics

Some inspiration for the potential programming language designer

For anyone thinking that lisp isn't abstract and self referential enough

First draft level.

Start with an "everything is a list" approach from lisp/scheme.

First order propositional logic can be defined in terms of the symbols $(,)$, $\Longrightarrow$, $\perp$ and the propositions $p_1, p_2, \cdots$. When you are trying to prove something about propositional logic, the fewer symbols to deal with the better. However, when you are trying to prove something in propositional logic, you want extra symbols like $\wedge$. This can be managed by considering $(a \wedge b)$ as syntactic shorthand for $(a \Longrightarrow (b \Longrightarrow \perp)) \Longrightarrow \perp$. Lisp style macros are good for this.

We can consider programs as formulas in some first order theory.

Consider the program (+ 5 (* 2 3)) The interpreter can syntactically modify it into the simpler program (+ 5 6) and then into 11.

A more complex example. (sum (map '(1 2 3) (lambda (x) (* x x)))) goes to (sum '((* 1 1) (* 2 2) (* 3 3))) then to (sum '(1 4 9)) and finally to 14.

Considered like this, the interpreter is automatically generating a proof that the program is equal to some value. Its automatically simplifying the program until it gets its answer. Note that it doesn't harness the full power of first order arithmetic, its missing predicates like $(\forall n \in N)$. It can also get stuck in infinite loops.

In the general view, there are a finite number of transformations that can be applied, and you want a sequence of transformations that leads from one tree to another. These transformations are the **atomic tactics**. (They are equivalent to axioms in some sense).

Example (forall (x N) (< x 7)) can be transformed by the general rule that for any $y \in$ N and function p : N → Bool, expressions of the form (forall (x N) (p x)) are equivalent to (and (p y) (forall (x N) (p x))). This gives (y=9) (and (< 9 7) (forall (x N) (< x 7))). This turns into (and False (forall (x N) (< x 7))) and then False.

Note that the only difference is that here the choice of local transformations is not uniquely determined. forall makes the abstract symbol x available in its interior in

much the same way that lisps "let" does. Here N the natural number type. This language would be typed at parse time, the type of an expression should depend only on the type of its sub-components. (A strict interpretation might have **less_reals** and **less_ints** distinct functions, one of type, but using the < symbol for both should work.) A type is just a variable of type type, so using a type that you haven't defined, and isn't builtin is just a special case of using an undefined variable. When parsing the code, use of any undefined variable fails syntactically. Types can be combined with union and struct as common in strongly typed programming languages. Standard category theory based type system.

This is where the idea of **tactics** comes in. (word definition) A **tactic** is a function that takes in an arbitrary formulae (and possibly some other stuff), and processes it and outputs a semantically identical formulae. (with runtime bounds on it doing so if need be, and the possibility of arbitrary programmer supplied hints)(it applies a series of syntactic transformations)

Here a function is any expression with free variables to substitute.

(lambda (x) (exists (y N) (= (* y 2) x))) is a perfectly good function, which returns

(exists (y N) (= (* y 2) 4)) when called on 4.

One builtin tactic would be evaluate, which evaluates expressions and finite loops, and ignores any predicate. If you just wrapped the rest of the code in an evaluate, an never used any other tactics, you would have a side effect free version of lisp. (or something like it)

Another tactic might be called example, which removes (∃x) by having the

programmer give a suitable x.

A toy example would be this function, which transforms ∃y : p(y) into

(p(x)) ∨ (∃y : p(y)) . Ie (exists (y N) (= (* y 2) 4)) into (using x=2)

(or (= (* 2 2) 4) (exists (y N) (= (* y 2) 4)))

(lambda (f x) (if (= (car f) 'exists) (list 'or (cddr f) (replace f (caadr f) x)) f))

Another could be minimize_metric_neighborhood this would take in some metric, eg number of sub-expressions, and try a bunch of other tactics to minimize it.

other tactics would be induction. deduction_thm ect.

The important point is that the programmer is free to design new tactics. The job of the compiler/interpreter is to ensure the tactics transform code in a syntactically valid manor, and to implement the built in tactics.

Note that you need to use a tactic to define new tactics.

Suppose you didn't have the deduction theorem and you wanted to define it

Suppose you provide a function that takes in a proof using the deduction theorem, and outputs a proof not using the deduction theorem, without proving that this function always outputs a valid proof. All you have is a macro, a convenient programmer shortcut. Whenever you use the deduction theorem macro, the proof is expanded out behind the scenes. Convenient, but not a new tactic.

Suppose you prove that "any theorem that can be proved using the deduction theorem can also be proved without it", not necessarily in a constructive manor. Then any time you want to use the deduction theorem, the program can use that tactic without expanding it out in this specific case.

To avoid Lobian obstacles to do with self trusting proof systems, all tactics must have a rank, which is an ordinal.

Suppose a tactic A is defined as an arbitrary function from an expression e and a hint h to an output s.

You validate A by proving (using tactics $t_1, t_2 \cdots t_n$) that

$$\forall e : \exists i \in N : \exists r_1, r_2, \cdots r_i \in Q : \exists h_1, h_2, \cdots h_i \in \text{ hints} : A(e,h) = r_1(r_2(\cdots r_i(e,h_i), \cdots, h_2), h_1)$$

(You needn't calculate r_i or h_i explicitly, just show that they must exist.)

Where $Q \subset$ tactics is a set such that $\sup_{q \in Q}(\text{rank}(q) + 1) = \alpha$ where α is some ordinal.

Then the rank of A must be chosen such that $\text{rank}(A) \geq \alpha$ and

$\forall j \leq n : \text{rank}(A) \geq \text{rank}(t_j)$. Most of the time, these will be small finite ordinals that could be filled in automatically.

Note that any proof that uses tactics of rank at most α is a valid proof in the language of $ZFC + \alpha$. The basic language would have a ZFC set type, (you need it for the ordinals), but if it wasn't there, you should be able to define it by declaring a bunch of atomic tactics).

Questions to discuss in comments

1) Is it a good real world strategy to design new programming languages more suitable for AI?

2) Are there any features a good AI language might want other than an excessive amount of self reference and abstract mathematicallity.

3) Any more suggestions or ambiguities related to the programming language outlined above?

4) If a programing language like this already exists, let me know.

Is AI safety doomed in the long term?

Are there any measures that humanity can put in place to control a vastly (and increasingly) more intelligent race?

On the basis that humans determine the fate of other species on the planet, I cannot find any reasons for believing that a lesser intelligence can control a greater intelligence.
Which leads me to think that AI safety is at most about controlling the development of AI until it makes, and can implement, its own decisions about the fate of humanity.

Is this a common stance and I am naively catching up?
Or what are the counter arguments?

alternative history: what if Bayes rule had never been discovered?

In trying to understand how Bayesian probability is used, I'm curious to know what wouldn't have been possible without it. how important was it in the course of human discovery, and in turn, how it effected history.

I don't demand rigorous answers, feel free to speculate and throw possibilities as you like.

Bonus question: if Bayes didn't discover it, when would it have to be discovered? (full speculation mode)

Dishonest Update Reporting

Related to: Asymmetric Justice, Privacy, Blackmail

Previously (Paul Christiano): Epistemic Incentives and Sluggish Updating

The starting context here is the problem of what Paul calls sluggish updating. Bob is asked to predict the probability of a recession this summer. He said 75% in January, and how believes 50% in February. What to do? Paul sees Bob as thinking roughly this:

> If I stick to my guns with 75%, then I still have a 50-50 chance of looking smarter than Alice when a recession occurs. If I waffle and say 50%, then I won't get any credit even if my initial prediction was good. Of course if I stick with 75% now and only go down to 50% later then I'll get dinged for making a bad prediction right now—but that's little worse than what people will think of me immediately if I waffle.

Paul concludes that this is likely:

> Bob's optimal strategy depends on exactly how people are evaluating him. If they care exclusively about evaluating his performance in January then he should always stick with his original guess of 75%. If they care exclusively about evaluating his performance in February then he should go straight to 50%. In the more realistic case where they care about both, his optimal strategy is somewhere in between. He might update to 70% this week.
>
> This results in a pattern of "sluggish" updating in a predictable direction: once I see Bob adjust his probability from 75% down to 70%, I expect that his "real" estimate is lower still. In expectation, his probability is going to keep going down in subsequent months. (Though it's not a sure thing—the whole point of Bob's behavior is to hold out hope that his original estimate will turn out to be reasonable and he can save face.)

This isn't 'sluggish' updating, of the type we talk about when we discuss the Aumann Agreement Theorem and its claim that rational parties can't agree to disagree. It's dishonest update reporting. As Paul says, explicitly.

> I think this kind of sluggish updating is quite common—if I see Bob assign 70% probability to something and Alice assign 50% probability, I expect their probabilities to gradually inch towards one another rather than making a big jump. (If Alice and Bob were epistemically rational and honest, their probabilities would immediately take big enough jumps that we wouldn't be able to predict in advance who will end up with the higher number. Needless to say, this is not what happens!)
>
> Unfortunately, I think that sluggish updating isn't even the worst case for humans. It's quite common for Bob to double down with his 75%, only changing his mind at the last defensible moment. This is less easily noticed, but is even more epistemically costly.

When Paul speaks of Bob's 'optimal strategy' he does not include a cost to lying, or a cost to others getting inaccurate information.

This is a world where all one cares about is how one is evaluated, and lying and deceiving others is free as long as you're not caught. You'll get *exactly* what you incentivize.

What that definitely *won't* get you are a lot more than just accurate probability estimates.

The only way to get accurate probability estimates from Bob-who-is-happy-to-strategically-lie is to use a mathematical formula to reward Bob based on his log likelihood score. Or to have Bob bet in a prediction market, or another similar robust method. And then use that as the *entirety* of how one evaluates Bob. If human judgment is allowed in the process, the value of that will overwhelm any desire on Bob's part to be precise or properly update.

Since Bob is almost certainly in a human context where humans are evaluating him based on human judgments, that means all is mostly lost.

As Paul notes, *consistency* is crucial in how one is evaluated. Even bigger is *avoiding mistakes.*

Given the asymmetric justice of punishing mistakes and inconsistency that can be proven and identified, the strategic actor must seek cognitive privacy. The more others know about the path of your beliefs, the easier it will be for them to spot an inconsistency or a mistake. It's hard enough to give a reasonable answer once, but updating in a way that never can be shown to have *ever* made a mistake or been inconstant? Impossible.

A mistake or inconsistency are the *bad things* one must avoid getting docked points for.

Thus, Bob's full strategy, in addition to choosing probabilities that sound best and give the best cost/benefit payoffs in human intuitive evaluations of performance, is to *avoid making any clear statements of any kind.* When he must do so, he will do his best to be able to deny having done so. Bob will seek to *destroy the historical record* of his predictions and statements, and their path. And also *prevent the creation of any common knowledge, at all.* Any knowledge of the past situation, or the present outcome, could be shown to not be consistent with what Bob said, or what we believe Bob said, or what we think Bob implied. And so on.

Bob's *optimal strategy* is *full anti-epistemology.* He is opposed to knowledge.

In that context, Paul's suggested solutions seem highly unlikely to work.

His first suggestion is to *exclude information* – to judge Bob only by the aggregation of all of Bob's predictions, and ignore any changes. Not only does this throw away vital information, it also isn't realistic. Even if it was realistic for some people, others would still punish Bob for updating.

Paul's second suggestion is to make predictions about others' belief changes, which he himself notes 'literally wouldn't work.' And that it is 'a recipe for epistemic catastrophe.' The whole thing is convoluted and unnatural at best.

Paul's third and final suggestion is social disapproval of sluggish updating. As he notes, this twists social incentives potentially in good ways but likely in ways that make things worse:

Having noticed that sluggish updating is a thing, it's tempting to respond by just penalizing people when they seem to update sluggishly. I think that's a problematic response:

- I think the rational reaction to norms against sluggish updating may often be no updating at all, which is much worse.
- In general combating non-epistemic incentives with other non-epistemic incentives seems like digging yourself into a hole, and can only work if you balance everything perfectly. It feels much safer to just try to remove the non-epistemic incentives that were causing the problem in the first place.
- Sluggish updating isn't easy to detect in any given case. For example, suppose that Bob expects an event to happen, and if it does he expects to get a positive sign on any given day with 1% probability. Then if the event doesn't happen his probability will decay exponentially towards zero, falling in half every ~70 days. This will look like sluggish updating.

Bob already isn't excited about updating. He'd prefer to not update at all. He's upset about having had to give that 75% answer, because now if there's new information (including others' opinions) he can't keep saying 'probably' and has to give a new number, again giving others information to use as ammunition against him.

The reason he updated visibly, at all, was that not updating would have been inconsistent or otherwise punished. Punish updates for being *too small* on top of *already* looking bad for changing at all, and the chance you get the incentives right here are almost zero. Bob will game the system, one way or another. And now, you won't know *how* Bob is doing it. Before, you could know that Bob moving from 75% to 70% meant going to something lower, perhaps 50%. Predictable bad calibration is much easier to fix. Twist things into knots and there's no way to tell.

Meanwhile, Bob is going to reliably get evaluated as smarter and more capable than Alice, who for reasons of principle is going around reporting her probability estimates accurately. Those observing might even punish Alice further, as someone who does not know how the game is played, and would be a poor ally.

The best we can do, under such circumstances, if we want insight from Bob, is to do our best to make Bob believe we will reward him for updating correctly and reporting that update honestly, then consider Bob's incentives, biases and instincts, and attempt as best we can to back out what Bob actually believes.

As Paul notes, we can try to combat non-epistemic incentives with equal and opposite other non-epistemic incentives, but going deep on that generally only makes things more complex and rewards more attention to our procedures and how to trick us, giving Bob an even bigger advantage over Alice.

A last-ditch effort would be to give Bob sufficient skin in the game. If Bob directly benefits enough from us having accurate models, Bob might report more accurately. But outside of very small groups, there isn't enough skin in the game to go around. And that still assumes Bob thinks the way for the group to succeed is to be honest and create accurate maps. Whereas most people like Bob do not think that is how winners behave. Certainly not with vague things that don't have direct physical consequences, like probability estimates.

What can be done about this?

Unless we care enough, very little. We lost early. We lost on the meta level. We didn't Play in Hard Mode.

We accepted that Bob was optimizing for how Bob was evaluated, rather than Bob optimizing for accuracy. But we didn't evaluate Bob on that basis. We didn't place the virtues of honesty and truth-seeking above the virtue of looking good sufficiently to make Bob's 'look good' procedure evolve into 'be honest and seek truth.' We didn't work to instill epistemic virtues in Bob, or select for Bobs with or seeking those virtues.

We didn't reform the local culture.

And we didn't fire Bob the moment we noticed.

Game over.

I once worked for a financial firm that made this priority clear. On the very first day. You need to always be ready to explain and work to improve your reasoning. If we catch you lying, *about anything at all, ever,* including a *probability estimate,* that's it. You're fired. Period.

It didn't solve all our problems. More subtle distortionary dynamics remained, and some evolved as reactions to the local virtues, as they always do. For these and other reasons, that I will not be getting into here or in the comments, it ended up not being a good place for me. Those topics are for another day.

But they sure as hell didn't have to worry about the likes of Bob.

Hierarchy and wings

This is a linkpost for http://benjaminrosshoffman.com/hierarchy-wings/

(Note for LessWrong: This is more overtly about partisan politics than the norm, but I think it's not *more* about that than The Two-Party Swindle, from the Sequences, and it proposes a structural model that doesn't require people to be as stupid.)

There are a few points I didn't make in my post on blame games because they seemed extraneous to the core point, which are still important enough to write down.

Hierarchy

The Hierarchy game is a zero-sum game in which people closer to the center expropriate from people farther from the center, and use some of those resources to perpetuate the power imbalances that enable the expropriation. Players that fail to submit to expropriation by higher-level players are punished by those more-powerful players, often through intermediaries. Players that fail to help members of their class expropriate from those beneath them are excluded from their class, and often coordinated against more overtly.

This game isn't inherently majoritarian, - instead, it allows smaller groups to stably expropriate from larger ones, because every player has a short-run incentive to go along with the arrangement. Feudalism is a simple example of the hierarchy game. Modern states almost always have some hierarchical arrangements, such as the police and military, and (less formally) economic class.

Political handedness

Around the time of the French revolution - a replacement of Feudal arrangements with Modern states - people started using terms like "left" and "right" to refer to political orientations. These terms are related to natural structural coalitions within a modern democratic state.

Political parties don't overtly promise to expropriate from 49% on behalf of an arbitrary 51%. This is probably in part because this would be correctly viewed as a proposal to massively increase expropriative activity relative to other activity, accelerating the rate of expropriation, which actually isn't in the majority's interests, and would quickly undermine the democratic paradigm without providing a replacement to enforce property claims. Instead, appropriation is opportunistic, and political coalitions seem to be oriented around natural power bases which could in principle replace deliberative democracy.

Right-wing

One natural sort of organization to orient around is the formal hierarchy with a monopoly on force - the military and police. The staffing needs of these organizations are substantial, especially in wartime (democracies perform well in wars in part

because of their ability to mobilize a large share of the population without destabilizing their internal political arrangement) so they already form a natural constituency.

The obvious advantage of control over these organizations is in the event of a civil war, control over the army and police would be a massive advantage. So, building a group identity around those things is a pretty plausible way to expropriate the country from the other half.

The "right wing" is the part of the political spectrum that most resembles or is most naturally allied with this coordination strategy. Generally, if there's an identifiable majority group (ethnic, cultural, religious, etc.), the hierarchy of violence will perceive members of that group as more "central" and want to help them expropriate from minority groups more than vice versa, insofar as gaps in the rule of law allow this. People rewarded by the existing credit-allocation, the "upper classes," will also tend to favor and be favored by this arrangement.

Left-wing

The "left wing" is the natural complement to this strategy: a political "big tent" made up of all the noncentral groups. Such a coalition has a structural advantage as long as democratic institutions persist, since any new group (e.g. immigrants) that isn't part of the majority is a natural member of the "left wing" coalition. Such groups also tend to seek control of, and expropriate resources through, the parts of the state that are responsible for information processing, investment, and resource allocation rather than the administration of violence. In short, the bureaucracies and those who staff them.

Related: Nightmare of the Perfectly Principled, Arseholes, considered as a strategic resource