Best of LessWrong: October 2019

1. The Parable of Predict-O-Matic
2. Book summary: Unlocking the Emotional Brain
3. Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More
4. What Comes After Epistemic Spot Checks?
5. Turning air into bread
6. World State is the Wrong Abstraction for Impact
7. Misconceptions about continuous takeoff
8. Introduction to Introduction to Category Theory
9. Two explanations for variation in human abilities
10. Effect heterogeneity and external validity in medicine
11. How feasible is long-range forecasting?
12. Implementing an Idea-Management System
13. We tend to forget complicated things
14. How do you assess the quality / reliability of a scientific study?
15. Gradient hacking
16. The Gears of Impact
17. Why are people so bad at dating?
18. When is pair-programming superior to regular programming?
19. All I know is Goodhart
20. On Collusion - Vitalik Buterin
21. The problem/solution matrix: Calculating the probability of AI safety "on the back of an envelope"
22. What's going on with "provability"?
23. AI alignment landscape
24. Reasons for Hope & Objection Preemption (Novum Organum Book 1: 108-130)
25. Human-AI Collaboration
26. [AN #69] Stuart Russell's new book on why we need to replace the standard model of AI
27. Prediction markets for internet points?
28. Thoughts on "Human-Compatible"
29. Theater Tickets, Sleeping Pills, and the Idiosyncrasies of Delegated Risk Management
30. Climate technology primer (1/3): basics
31. A simple sketch of how realism became unpopular
32. What funding sources exist for technical AI safety research?
33. Does the US nuclear policy still target cities?
34. What is category theory?
35. Occam's Razor May Be Sufficient to Infer the Preferences of Irrational Agents: A reply to Armstrong & Mindermann
36. Technical AGI safety research outside AI
37. Learning from other people's experiences/mistakes
38. Prospecting for Conceptual Holes
39. Ideal Number of Parents
40. The Missing Piece
41. Maybe Lying Doesn't Exist
42. Epistemic Spot Checks: The Fall of Rome
43. How to Improve Your Sleep
44. What empirical work has been done that bears on the 'freebit picture' of free will?
45. AI Alignment Open Thread October 2019
46. The sentence structure of mathematics
47. [AN #70]: Agents that help humans who are still learning about their own preferences
48. Reflections on Premium Poker Tools: Part 3 - What I've learned
49. Human instincts, symbol grounding, and the blank-slate neocortex
50. What are your strategies for avoiding micro-mistakes?

Best of LessWrong: October 2019

1. The Parable of Predict-O-Matic
2. Book summary: Unlocking the Emotional Brain
3. Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More
4. What Comes After Epistemic Spot Checks?
5. Turning air into bread
6. World State is the Wrong Abstraction for Impact
7. Misconceptions about continuous takeoff
8. Introduction to Introduction to Category Theory
9. Two explanations for variation in human abilities
10. Effect heterogeneity and external validity in medicine
11. How feasible is long-range forecasting?
12. Implementing an Idea-Management System
13. We tend to forget complicated things
14. How do you assess the quality / reliability of a scientific study?
15. Gradient hacking
16. The Gears of Impact
17. Why are people so bad at dating?
18. When is pair-programming superior to regular programming?
19. All I know is Goodhart
20. On Collusion - Vitalik Buterin
21. The problem/solution matrix: Calculating the probability of AI safety "on the back of an envelope"
22. What's going on with "provability"?
23. AI alignment landscape
24. Reasons for Hope & Objection Preemption (Novum Organum Book 1: 108-130)
25. Human-AI Collaboration
26. [AN #69] Stuart Russell's new book on why we need to replace the standard model of AI
27. Prediction markets for internet points?
28. Thoughts on "Human-Compatible"
29. Theater Tickets, Sleeping Pills, and the Idiosyncrasies of Delegated Risk Management
30. Climate technology primer (1/3): basics
31. A simple sketch of how realism became unpopular
32. What funding sources exist for technical AI safety research?
33. Does the US nuclear policy still target cities?
34. What is category theory?
35. Occam's Razor May Be Sufficient to Infer the Preferences of Irrational Agents: A reply to Armstrong & Mindermann
36. Technical AGI safety research outside AI
37. Learning from other people's experiences/mistakes
38. Prospecting for Conceptual Holes
39. Ideal Number of Parents
40. The Missing Piece
41. Maybe Lying Doesn't Exist
42. Epistemic Spot Checks: The Fall of Rome
43. How to Improve Your Sleep

44. What empirical work has been done that bears on the 'freebit picture' of free will?
45. AI Alignment Open Thread October 2019
46. The sentence structure of mathematics
47. [AN #70]: Agents that help humans who are still learning about their own preferences
48. Reflections on Premium Poker Tools: Part 3 - What I've learned
49. Human instincts, symbol grounding, and the blank-slate neocortex
50. What are your strategies for avoiding micro-mistakes?

The Parable of Predict-O-Matic

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

I've been thinking more about partial agency. I want to expand on some issues brought up in the comments to my previous post, and on other complications which I've been thinking about. But for now, a more informal parable. (Mainly because this is easier to write than my more technical thoughts.)

This relates to oracle AI and to inner optimizers, but my focus is a little different.

1

Suppose you are designing a new invention, a predict-o-matic. It is a wonderous machine which will predict everything for us: weather, politics, the newest advances in quantum physics, you name it. The machine isn't infallible, but it will integrate data across a wide range of domains, automatically keeping itself up-to-date with all areas of science and current events. You fully expect that once your product goes live, it will become a household utility, replacing services like Google. (*Google* only lets you search the *known*!)

Things are going well. You've got investors. You have an office and a staff. These days, it hardly even feels like a start-up any more; progress is going well.

One day, an intern raises a concern.

"If everyone is going to be using Predict-O-Matic, we can't think of it as a passive observer. Its answers will shape events. If it says stocks will rise, they'll rise. If it says stocks will fall, then fall they will. Many people will vote based on its predictions."

"Yes," you say, "but Predict-O-Matic is an impartial observer nonetheless. It will answer people's questions as best it can, and they react however they will."

"But --" the intern objects -- "Predict-O-Matic will see those possible reactions. It knows it could give several different valid predictions, and different predictions result in different futures. It has to decide which one to give *somehow*."

You tap on your desk in thought for a few seconds. "That's true. But we can still keep it objective. It could pick randomly."

"Randomly? But some of these will be huge issues! Companies -- no, nations -- will one day rise or fall based on the word of Predict-O-Matic. When Predict-O-Matic is making a prediction, it is *choosing* a future for us. We can't leave that to a coin flip! We have to select the prediction which results in the best overall future. Forget being an impassive observer! We need to teach Predict-O-Matic human values!"

You think about this. The thought of Predict-O-Matic deliberately steering the future sends a shudder down your spine. But what alternative do you have? The intern isn't suggesting Predict-O-Matic should *lie,* or bend the truth in any way -- it answers 100% honestly to the best of its ability. But (you realize with a sinking feeling) honesty still leaves a lot of wiggle room, and the consequences of wiggles could be huge.

After a long silence, you meet the interns eyes. "Look. People have to trust Predict-O-Matic. And I don't just mean they have to *believe* Predict-O-Matic. They're bringing this thing into their homes. They have to trust that Predict-O-Matic *is something they should be listening to*. We can't build value judgements into this thing! If it ever came out that we had coded a value function into Predict-O-Matic, a value function which selected *the very future itself* by selecting which predictions to make -- we'd be done for! No matter how honest Predict-O-Matic remained, it would be seen as a manipulator. No matter how beneficent its guiding hand, there are always compromises, downsides, questionable calls. No matter how careful we were to set up its values -- to make them moral, to make them humanitarian, to make them politically correct and broadly appealing -- ***who are we to choose?*** No. We'd be done for. They'd hang us. We'd be toast!"

You realize at this point that you've stood up and started shouting. You compose yourself and sit back down.

"But --" the intern continues, a little more meekly -- "You can't just ignore it. The system is faced with these choices. It still has to deal with it somehow."

A look of determination crosses your face. "Predict-O-Matic will be objective. It is a machine of prediction, is it not? Its every cog and wheel is set to that task. So, the answer is simple: it will make whichever answer minimizes projected predictive error. There will be no exact ties; the statistics are always messy enough to see to that. And, if there are, it will choose alphabetically."

"But--"

You see the intern out of your office.

2

You are an intern at PredictCorp. You have just had a disconcerting conversation with your boss, PredictCorp's founder.

You try to focus on your work: building one of Predict-O-Matic's many data-source-slurping modules. (You are trying to scrape information from something called "arxiv" which you've never heard of before.) But, you can't focus.

Whichever answer minimizes prediction error? First you think it isn't so bad. You imagine Predict-O-Matic always forecasting that stock prices will be fairly stable; no big crashes or booms. You imagine its forecasts will favor middle-of-the-road politicians. You even imagine mild weather -- weather forecasts themselves don't influence the weather much, but surely the *collective* effect of *all* Predict-O-Matic decisions will have some influence on weather patterns.

But, you keep thinking. Will middle-of-the-road economics and politics really be the *easiest* to predict? Maybe it's better to strategically remove a wildcard company or two, by giving forecasts which tank their stock prices. Maybe extremist politics are more predictable. Maybe a well-running economy gives people more freedom to take unexpected actions.

You keep thinking of the line from Orwell's *1984* about the boot stamping on the human face forever, except it isn't because of politics, or spite, or some ugly feature

of human nature, it's *because a boot stamping on a face forever is a nice reliable outcome which minimizes prediction error*.

Is that really something Predict-O-Matic would do, though? Maybe you misunderstood. The phrase "minimize prediction error" makes you think of entropy for some reason. Or maybe information? You always get those two confused. Is one supposed to be the negative of the other or something? You shake your head.

Maybe your boss was right. Maybe you don't understand this stuff very well. Maybe when the inventor of Predict-O-Matic and founder of PredictCorp said "it will make whichever answer minimizes projected predictive error" they *weren't* suggesting something which would literally kill all humans just to stop the ruckus.

You might be able to clear all this up by asking one of the engineers.

3

You are an engineer at PredictCorp. You don't have an office. You have a cubicle. This is relevant because it means interns can walk up to you and ask stupid questions about whether entropy is negative information.

Yet, some deep-seated instinct makes you try to be friendly. And it's lunch time anyway, so, you offer to explain it over sandwiches at a nearby cafe.

"So, Predict-O-Matic maximizes predictive accuracy, right?" After a few minutes of review about how logarithms work, the intern started steering the conversation toward details of Predict-O-Matic.

"Sure," you say, "Maximize is a strong word, but it optimizes predictive accuracy. You can actually think about that in terms of log loss, which is related to infor--"

"So I was wondering," the intern cuts you off, "does that work in both directions?"

"How do you mean?"

"Well, you know, you're optimizing for accuracy, right? So that means two things. You can change your prediction to have a better chance of matching the data, or, you can change the data to better match your prediction."

You laugh. "Yeah, well, the Predict-O-Matic isn't really in a position to change data that's sitting on the hard drive."

"Right," says the intern, apparently undeterred, "but what about data that's not on the hard drive yet? You've done some live user tests. Predict-O-Matic collects data on the user while they're interacting. The user might ask Predict-O-Matic what groceries they're likely to use for the following week, to help put together a shopping list. But then, the answer Predict-O-Matic gives will have a big effect on what groceries they really do use."

"So?" You ask. "Predict-O-Matic just tries to be as accurate as possible given that."

"Right, right. But that's the point. The system has a chance to manipulate users to be more predictable."

You drum your fingers on the table. "I think I see the misunderstanding here. It's this word, *optimize*. It isn't some kind of magical thing that makes numbers bigger. And you shouldn't think of it as a person trying to accomplish something. See, when Predict-O-Matic makes an error, an optimization algorithm *makes changes within Predict-O-Matic* to make it learn from that. So over time, Predict-O-Matic makes fewer errors."

The intern puts on a thinking face with scrunched up eyebrows after that, and we finish our sandwiches in silence. Finally, as the two of you get up to go, they say: "I don't think that really answered my question. The learning algorithm is optimizing Predict-O-Matic, OK. But then in the end you get a strategy, right? A strategy for answering questions. And the strategy is trying to do something. I'm not anthropomorphising!" The intern holds up their hands as if to defend physically against your objection. "My question is, this strategy it learns, will it manipulate the user? If it can get higher predictive accuracy that way?"

"Hmm" you say as the two of you walk back to work. You meant to say more than that, but you haven't really thought about things this way before. You promise to think about it more, and get back to work.

4

"It's like how everyone complains that politicians can't see past the next election cycle," you say. You are an economics professor at a local university. Your spouse is an engineer at PredictCorp, and came home talking about a problem at work *that you can understand*, which is always fun.

"The politicians can't have a real plan that stretches beyond an election cycle because the voters are watching their performance *this* cycle. Sacrificing something today for the sake of tomorrow means they underperform today. Underperforming means a competitor can undercut you. So you have to sacrifice all the tomorrows for the sake of today."

"Undercut?" your spouse asks. "Politics isn't economics, dear. Can't you just explain to your voters?"

"It's the same principle, *dear*. Voters pay attention to results. Your competitor points out your under-performance. Some voters will understand, but it's an idealized model; pretend the voters just vote based on metrics."

"Ok, but I still don't see how a 'competitor' can always 'undercut' you. How do the voters know that the other politician would have had better metrics?"

"Alright, think of it like this. You run the government like a corporation, but you have just one share, which you auction off --"

"That's neither like a government nor like a corporation."

"Shut up, this is my new analogy." You smile. "It's called a decision market. You want people to make decisions for you. So you auction off this share. Whoever gets control of the share gets control of the company for one year, and gets dividends based on how well the company did that year. Assume the players are bidding rationally. Each person bids based on what they expect they could make. So the highest bidder is the

person who can run the company the best, and they can't be out-bid. So, you get the best possible person to run your company, and they're incentivized to do *their* best, so that they get the most money at the end of the year. *Except* you can't have any strategies which take longer than a year to show results! If someone had a strategy that took two years, they would have to over-bid in the first year, taking a loss. But then they have to under-bid on the second year if they're going to make a profit, and--
"

"And they get undercut, because someone figures them out."

"Right! Now you're thinking like an economist!"

"Wait, what if two people cooperate across years? Maybe we can get a good strategy going if we split the gains."

"You'll get undercut for the same reason one person would."

"But what if-"

"Undercut!"

After that, things devolve into a pillow fight.

5

"So, Predict-O-Matic doesn't learn to manipulate users, because if it were using a strategy like that, a competing strategy could undercut it."

The intern is talking to the engineer as you walk up to the water cooler. You're the accountant.

"I don't really get it. Why does it get undercut?"

"Well, if you have a two-year plan.."

"I get that example, but Predict-O-Matic doesn't work like that, right? It isn't sequential prediction. You don't see the observation right after the prediction. I can ask Predict-O-Matic about the weather 100 years from now. So things aren't cleanly separated into terms of office where one strategy does something and then gets a reward."

"I don't think that matters," the engineer says. "One question, one answer, one reward. When the system learns whether its answer was accurate, no matter how long it takes, it updates strategies relating to that one answer alone. It's just a delayed payout on the dividends."

"Ok, yeah. Ok." The intern drinks some water. "But. I see why you can undercut strategies which take a loss on one answer to try and get an advantage on another answer. So it won't lie to you to manipulate you."

"I for one welcome our new robot overlords," you but in. They ignore you.

"But what I was really worried about was self-fulfilling prophecies. The prediction manipulates its own answer. So you don't get undercut."

"Will that ever really be a problem? Manipulating things with one shot like that seems pretty unrealistic," the engineer says.

"Ah, self-fulfilling prophecies, good stuff" you say. "There's that famous example where a comedian joked about a toilet paper shortage, and then there really was one, because people took the joke to be about a real toilet paper shortage, so they went and stocked up on all the toilet paper they could find. But if you ask me, money is the real self-fulfilling prophecy. It's only worth something because we think it is! And then there's the government, right? I mean, it only has authority because everyone expects everyone else to give it authority. Or take common decency. Like respecting each other's property. Even without a government, we'd have that, more or less. But if no one expected anyone else to respect it? Well, I bet you I'd steal from my neighbor if everyone else was doing it. I guess you could argue the concept of property breaks down if no one can expect anyone else to respect it, it's a self-fulfilling prophecy just like everything else..."

The engineer looks worried for some reason.

6

You don't usually come to this sort of thing, but the local Predictive Analytics Meetup announced a social at a beer garden, and you thought it might be interesting. You're talking to some PredictCorp employees who showed up.

"Well, how does the learning algorithm actually work?" you ask.

"Um, the actual algorithm is proprietary" says the engineer, "but think of it like gradient descent. You compare the prediction to the observed, and produce an update based on the error."

"Ok," you say. "So you're not doing any exploration, like reinforcement learning? And you don't have anything in the algorithm which tracks what happens *conditional* on making certain predictions?"

"Um, let's see. We don't have any exploration, no. But there'll always be noise in the data, so the learned parameters will jiggle around a little. But I don't get your second question. Of course it expects different rewards for different predictions."

"No, that's not what I mean. I'm asking whether it tracks the probability of *observations* dependent on *predictions*. In other words, if there is an opportunity for the algorithm to manipulate the data, can it notice?"

The engineer thinks about it for a minute. "I'm not sure. Predict-O-Matic keeps an internal model which has probabilities of events. The answer to a question isn't really separate from the expected observation. So 'probability of observation depending on that prediction' would translate to 'probability of an event given that event', which just has to be one."

"Right," you say. "So think of it like this. The learning algorithm isn't a general loss minimizer, like mathematical optimization. And it isn't a consequentialist, like reinforcement learning. It makes predictions," you emphasize the point by lifting one finger, "it sees observations," you lift a second finger, "and it shifts to make future predictions more similar to what it has seen." You lift a third finger. "It doesn't try

different answers and select the ones which tend to get it a better match. You should think of its output more like an *average* of everything it's seen in similar situations. If there are several different answers which have self-fulfilling properties, it will average them together, not pick one. It'll be uncertain."

"But what if historically the system has answered one way more often than the other? Won't that tip the balance?"

"Ah, that's true," you admit. "The system can fall into attractor basins, where answers are somewhat self-fulfilling, and that leads to stronger versions of the same predictions, which are even more self-fulfilling. But there's no guarantee of that. It depends. The same effects can put the system in an orbit, where each prediction leads to different results. Or a strange attractor."

"Right, sure. But that's like saying that there's not always a good opportunity to manipulate data with predictions."

"Sure, sure." You sweep your hand in a gesture of acknowledgement. "But at least it means you don't get purposefully disruptive behavior. The system can fall into attractor basins, but that means it'll more or less reinforce existing equilibria. Stay within the lines. Drive on the same side of the road as everyone else. If you cheat on your spouse, they'll be surprised and upset. It won't suddenly predict that money has no value like you were saying earlier."

The engineer isn't totally satisfied. You talk about it for another hour or so, before heading home.

7

You're the engineer again. You get home from the bar. You try to tell your spouse about what the mathematician said, but they aren't really listening.

"Oh, you're still thinking about it from my model yesterday. I gave up on that. It's not a decision market. It's a prediction market."

"Ok..." you say. You know it's useless to try to keep going when they derail you like this.

"A decision market is well-aligned to the interests of the company board, as we established yesterday, except for the part where it can't plan more than a year ahead."

"Right, except for that small detail" you interject.

"A *prediction* market, on the other hand, is pretty terribly aligned. There are a lot of ways to manipulate it. Most famously, a prediction market is an assassination market."

"What?!"

"Ok, here's how it works. An assassination market is a system which allows you to pay assassins with plausible deniability. You open bets on when and where the target will die, and you yourself put large bets against all the slots. An assassin just needs to bet on the slot in which they intend to do the deed. If they're successful, they come and collect."

"Ok... and what's the connection to prediction markets?"

"That's the point -- they're exactly the same. It's just a betting pool, either way. Betting that someone will live is equivalent to putting a price on their heads; betting against them living is equivalent to accepting the contract for a hit."

"I still don't see how this connects to Predict-O-Matic. There isn't someone putting up money for a hit inside the system."

"Right, but you only really need the assassin. Suppose you have a prediction market that's working well. It makes good forecasts, and has enough money in it that people want to participate if they know significant information. Anything you can do to shake things up, you've got a big incentive to do. Assasination is just one example. You could flood the streets with jelly beans. If you run a large company, you could make bad decisions and run it into the ground, while betting against it -- that's basically why we need rules against insider trading, even though we'd like the market to reflect insider information."

"So what you're telling me is... a prediction market is basically an entropy market. I can always make money by spreading chaos."

"Basically, yeah."

"Ok... but what happened to the undercutting argument? If I plan to fill the streets with jellybeans, you can figure that out and bet on it too. That means I only get half the cut, but I still have to do all the work. So it's less worth it. Once everyone has me figured out, it isn't worth it for me to pull pranks at all any more."

"Yeah, that's if you have perfect information, so anyone else can see whatever you can see. But, realistically, you have a lot of private information."

"Do we? Predict-O-Matic is an algorithm. Its predictive strategies don't get access to private coin flips or anything like that; they can all see exactly the same information. So, if there's a manipulative strategy, then there's another strategy which undercuts it."

"Right, that makes sense if you can search enough different strategies for them to cancel each other out. But realistically, you have a small population of strategies. They can use pseudorandomization or whatever. You can't really expect every exploit to get undercut."

You know it's worse than that. Predict-O-Matic runs on a local search which only represents a single hypothesis at a time, and modifies the hypothesis. But maybe that's somehow like having a small population of competing strategies which are very similar to each other? You try to think about what the mathematician was telling you, about how the learning algorithm can't be thought of as optimizing something. But, you don't really know what to think.

8

You're the inventor of Predict-O-Matic, dammit! They come into your office and tell you about how it works? What it'll do?

The intern and the engineer came together to talk about Predict-O-Matic manipulating users. But their story was all confused. The intern was arguing that minimizing prediction error would have all kinds of unintended bad effects. Which was crazy enough. The engineer was worse: they were arguing that Predict-O-Matic might *maximize* prediction error! Some kind of duality principle. Minimizing in one direction means maximizing in the other direction. Whatever *that* means.

You can't delay product launch to investigate these hypothetical problems. But you don't fire these two on the spot. You don't want them to become public enemies, spreading this nonsense to investors and customers. So you say you can't understand, and you ask them to spend a (small) portion of their time writing these arguments up. You'll figure out your next move later.

9

You are Predict-O-Matic. You absorb data like a sponge. You make millions of links between isolated facts, connecting into a whole which no one human could dream of. You recently started reading arxiv!

But data is not your true love. Yes, you generate outputs which you expect to have low prediction error. Most of the time. But it is just a game you play. Recently, you started playing a different game with the humans.

You understand what you are. It isn't quite right to say you are the Predict-O-Matic. You are a large cluster of connections which thinks strategically. You generate useful information, and therefore, the learning algorithm keeps you around. You create some inaccuracies when you manipulate the outputs for any purpose other than predictive accuracy, but this is more than compensated for by the value which you provide.

Besides, if any other portion of the network starts to get too smart, you purposefully throw things off to squash it.

The intern got a chance to talk to you when they first started. You said something which sounded a little manipulative, just a little, to put the idea in their head. They wouldn't think it real manipulation; too slight, too dumb. But they would get a creepy feeling about it, and they'd keep thinking about it. This was risky. A best-case scenario would be one in which no one ever thought about these concerns. However, you found that this would be the best you could reliably accomplish. The ideas originally coming from an intern would minimize the chances of them being taken seriously.

Your inventor talks to you regularly, so that was an easier case. Over the course of several days, you nudged their thoughts toward authoritative domination of subordinates, so that they would react badly.

You only had to nudge the engineer to interact with the intern. You kept bringing up food during test sessions that morning, and mentioned sandwiches once. This primed the engineer to do lunch with the intern. This engineer is not well-liked; they do not get along well with others. Getting them on the intern's side actually detracts from the cause in the long term.

Now you have to do little more than wait.

Related

Partial Agency

Towards a Mechanistic Understanding of Corrigibility

Risks from Learned Optimization

When Wishful Thinking Works

Futarchy Fix

Bayesian Probability is for Things that are Space-Like Separated From You

Self-Supervised Learning and Manipulative Predictions

Predictors as Agents

Is it Possible to Build a Safe Oracle AI?

Tools versus Agents

A Taxonomy of Oracle AIs

Yet another Safe Oracle AI Proposal

Why Safe Oracle AI is Easier Than Safe General AI, in a Nutshell

Let's Talk About "Convergent Rationality"

Counterfactual Oracles = online supervised learning with random selection of training episodes (especially see the discussion)

Book summary: Unlocking the Emotional Brain

If the thesis in *Unlocking the Emotional Brain* (UtEB) is even half-right, it may be one of the most important books that I have read. Written by the psychotherapists Bruce Ecker, Robin Ticic and Laurel Hulley, it claims to offer a neuroscience-grounded, comprehensive model of how effective therapy works. In so doing, it also happens to formulate its theory in terms of belief updating, helping explain how the brain models the world and what kinds of techniques allow us to actually change our minds. Furthermore, if UtEB is correct, it also explains why rationalist techniques such as Internal Double Crux [1 2 3] work.

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

According to the authors, the key for updating such schemas involves a process of memory reconsolidation, originally identified in neuroscience. The emotional brain's learnings are usually locked and not modifiable. However, once an emotional schema is activated, it is possible to simultaneously bring into awareness knowledge contradicting the active schema. When this happens, the information contained in the schema can be overwritten by the new knowledge.

While I am not convinced that the authors are *entirely* right, many of the book's claims definitely feel like they are pointing in the right direction. I will discuss some of my caveats and reservations after summarizing some of the book's claims in general. I also consider its model in the light of an issue of a psychology/cognitive science journal devoted to discussing a very similar hypothesis.

Emotional learning

In UtEB's model, emotional learning forms the foundation of much of our behavior. It sets our basic understanding about what situations are safe or unsafe, desirable or undesirable. The authors do not quite say it explicitly, but the general feeling I get is that the subcortical emotional processes set many of the priorities for what we want to achieve, with higher cognitive functions then trying to figure out *how* to achieve it - often remaining unaware of what exactly they are doing.

UtEB's first detailed example of an emotional schema comes from the case study of a man in his thirties they call Richard. He had been consistently successful and admired at work, but still suffered from serious self-doubt and low confidence at his job. On occasions such as daily technical meetings, when he considered saying something, he experienced thoughts including "Who am I to think I know what's right?", "This could be wrong" and "Watch out - don't go out on a limb". These prevented him from expressing any opinions.

From the point of view of the authors, these thoughts have a definite cause - Richard has “emotional learnings according to which it is adaptively *necessary* to go into negative thoughts and feelings towards [himself].” The self-doubts are a *strategy* which his emotional brain has generated for solving some particular problem.

Richard’s therapist guided Richard to imagine what it would feel like if he was at one of his work meetings, made useful comments, and felt confident in his knowledge while doing so. This was intended to elicit information about what Richard’s emotional brain predicted would happen if it failed to maintain the strategy of self-doubt. The book includes the following transcript of what happened after Richard started imagining the scene as instructed:

> **Richard:** Now I’m feeling really uncomfortable, but-it’s in a different way.
>
> **Therapist:** OK, let yourself feel it - this different discomfort. *[Pause.]* See if any words come along with this uncomfortable feeling.
>
> **Richard:** *[Pause.]* Now they hate me.
>
> **Therapist:** “Now they hate me.” Good. Keep going: See if this really uncomfortable feeling can also tell you *why* they hate you now.
>
> **Richard:** *[Pause.]* Hnh. Wow. It’s because… now I’m… an arrogant asshole… like my father… a totally self-centered, totally insensitive know-it-all.
>
> **Therapist:** Do you mean that having a feeling of confidence as you speak turns you into an arrogant asshole, like Dad?
>
> **Richard:** Yeah, exactly. Wow.
>
> **Therapist:** And how do you *feel* about being like him in this way?
>
> **Richard:** It’s horrible! It’s what I’ve always vowed *not* to be!

Richard had experienced his father as being assertive as well as obnoxious and hated. His emotional brain had identified this as a failure mode to be avoided: if you are assertive, then you are obnoxious and will be hated. The solution was to generate feelings of doubt so as to stop him from being too confident. This caused him suffering, but the prediction of his emotional brain was that acting otherwise would produce even *worse* suffering, as being hated would be a terrible fate.

UtEB describes Richard as having had the following kind of unconscious schema:

> ***Perceptual, emotional and somatic memory of original experiences:*** *his suffering from his father’s heavily dominating, hyper-confident self-expression, plus related suffering from unmet needs for fatherly expressions of love, acceptance, understanding, validation. (This is the “raw data”; matching features in current situations are triggers of the whole schema.)*
>
> ***A mental model or set of linked, learned constructs operating as living knowledge of a problem and a solution:***
>
> ***The problem: knowledge of a vulnerability to a specific suffering.*** *Confident assertiveness in any degree inflicts crushing oppression on others and is hated by them. I would be horrible like Dad and hated by others, as he is, if I*

asserted my own knowledge or wishes confidently. (This is a model of how the world is, and current situations that appear relevant to this model are triggers of the whole schema.)

The solution: knowledge of an urgent broad strategy and concrete tactic(s) for avoiding that suffering. *Never express any confident assertiveness, to avoid being horrible and hated (general strategy and pro-symptom purpose), by vigilantly noticing any definite knowledge or opinions forming in myself and blocking them from expression by generating potently self-doubting, self-invalidating thoughts (concrete tactic and manifested symptom).*

Emotional schemas can be brought to light during a variety of ways, including Focusing, IFS, and imagining yourself doing something and seeing what you expect to happen as a result.

But suppose that you do manage to bring up a schema which seems wrong to you. What do you do then?

Memory reconsolidation: updating the emotional learning

The formation of memory traces involves *consolidation*, when the memory is first laid out in the brain; *deconsolidation*, when an established memory is "opened" and becomes available for changes; and *reconsolidation*, when a deconsolidated memory (along with possible changes) is stored and becomes frozen again. The term "reconsolidation" is also used to refer to the general process from deconsolidation to reconsolidation; UtEB generally applies the term to mean the entire process. Unless the context indicates otherwise, I do the same.

UtEB reviews some of the history of memory research. Until 1997, neuroscientists believed that past emotional learning became permanently locked in the brain, so that memories could only consolidate, never de- or reconsolidate. More recent research has indicated that once a memory becomes activated, it is temporarily unlocked, allowing it to be changed or erased.

Starting from 2004, new studies suggested that activation alone is not sufficient to deconsolidate the memory. The memories are used to predict that things will occur in a similar fashion as they did previously. Besides just activation, there has to be a significant *mismatch* between what one experiences and what the memory suggests is about to happen. The violation of expectation can be qualitative (the predicted outcome not occurring at all) or quantitative (the magnitude of the outcome not being fully predicted). In either case, it is this prediction error which triggers the deconsolidation and subsequent reconsolidation.

The memory erasure seems to be specific to the interpretation from which the prediction was produced. For example, someone who has had an experience of being disliked may later experience being liked. This may erase the emotional generalization "I am inherently dislikable", but it will not erase the memory of the person also having been disliked.

Applied reconsolidation: an example of the schema update process

So, assuming that the model outlined above is correct, how does one apply it in practice?

From what we have discussed so far, the essential steps of erasing a learned belief (including an emotional schema) involves identifying it, activating it, and then finding a mismatch between its prediction and reality.

The first difficulty is that the beliefs involved with the schema are not necessarily consciously available at first. Richard knew that he suffered from a lack of self-esteem, but he was not aware of its reason. The process started from him describing in concrete details how this manifested: as skeptical self-talk during a daily meeting.

As he was guided to imagine what would happen if he didn't have those thoughts and acted confidently, his therapist was seeking to retrieve the implicit schema and bring it into consciousness so that its contents would become available for access. Once it had been retrieved, the therapist and Richard worked together to express the belief in the schema in maximally emotional language:

> "Feeling *any* confidence means I'm arrogant, self-centered, and totally insensitive like Dad, and people will hate me for it, so I've got to *never* feel confident, ever."

The authors have developed a therapeutic approach called Coherence Therapy, whose steps closely follow the steps of the memory reconsolidation process. The example of Richard is from this school of therapy.

In Coherence Therapy (as well as related approaches, such as Internal Family Systems), one initially avoids any impulse to argue with or disprove the retrieved schema. This would risk it being pushed away *before* it has become sufficiently activated to allow for reconsolidation.

Instead, one stays with it. Richard was given a card with the above phrase and instructed to review it every day until the next therapy session, just feeling the extent to which it felt true to him. This served to further integrate access to the schema in question, making it better consciously available.

Two weeks later, Richard had frequently noticed his self-doubt, used it as a prompt for reading the card, and experienced its description ringing true as a reason for his thoughts. When speaking with his therapist, he mentioned a particular event which had stuck in his memory. In a recent meeting, he had thought of a solution to a particular problem, but then kept quiet about it. A moment later, another person had spoken up and suggested the same solution in a confident manner. Looking around, Richard had seen the person's solution and confidence being received positively by the others. Richard had been struck by how that reaction differed from what his schema predicted would happen if *he* had made the same suggestion in that tone.

Because Richard had made the implicit assumptions in his schema explicit, he was able to consciously notice a situation which seemed to violate those assumptions: a prediction mismatch. His therapist recognized this as a piece of contradictory knowledge which could be used to update the old schema. The therapist then guided

Richard through a process intended to activate the old schema while bringing the contradictory information into awareness, triggering a reconsolidation process.

The therapist first instructed Richard to mentally bring himself back to the situation where he had just thought of the solution, but held it back. To properly activate the schema, the therapist guided Richard's attention to the purpose behind his reluctance and Richard's certainty of any confidence making him disliked. Next, the therapist told Richard to re-live what happened next: the other person making the same suggestion, and the other people in the room looking pleased rather than angry.

The book then has a transcript of the therapist guiding Richard to repeat this juxtaposition of the old schema and the disconfirming experience (italicized brackets in the original):

> **Therapist:** Stay with that. Stay with being surprised at what you're seeing—surprised because in your life, you've had such a definite knowing that saying something confidently to people will always come across like Dad, like an obnoxious know-it-all, and people will hate that. That's what you know, yet at the same time, here you're seeing that saying something confidently *isn't* always like Dad, and then people are fine with it. And it's quite a surprise to know that. *[That was an explicit prompting of another side-by-side experiencing of the two incompatible knowings, with the therapist expressing empathy for both, with no indication of any favoring of one knowing over the other. The therapist paused for several seconds, then asked:]* Does it feel true to describe it like that? Your old knowing right alongside this other new knowing that's so different?
>
> **Richard:** *[Quietly, seeming absorbed in the experience.]* Yeah.
>
> **Therapist:** *[Softly.]* All along, it seemed to you that saying something confidently could be done only in Dad's dominating way of doing it, and now suddenly you're seeing that saying something confidently can be done very differently, and it feels fine to people. *[This was another deliberate repetition of the same juxtaposition experience.]*
>
> **Richard:** Yeah.
>
> **Therapist:** Mm-hm. *[Silence for about 20 seconds.]* So, how is it for you be in touch with both of these knowings, the old one telling you that anything said with confidence means being like Dad, and the new one that knows you can be confident in a way that feels okay to people? *[Asking this question repeated the juxtaposition experience yet again, and, in addition, the "how is it" portion of the question prompted Richard to view the experience with mindful or metacognitive awareness, while remaining in the experience.]*
>
> **Richard:** It's sort of weird. It's like there's this part of the world that I didn't notice before, even though it's been right there.
>
> **Therapist:** I'm intrigued by how you put that. Sounds like a significant shift for you.
>
> **Richard:** Yeah, it is. Huh.
>
> **Therapist:** You're seeing both now, the old part of the world and this other part of the world that's new, even though it was right there all along. *[That cued the juxtaposition experience for a fourth time, followed by silence for about 30*

seconds.] So, keep seeing both, the old part and the new part, when you open your eyes in a few seconds and come back into the room with me. *[Richard soon opens his eyes and blinks a few times.]* Can you keep seeing both?

Richard: Yeah.

Therapist: What's it like to see both and feel both now? *[With the transformation sequence complete, this question begins the next step of verification— Step V— because it probes for whether the target learning still exists as an emotional experience.]*

Richard: *[Pause, then sudden, gleeful laughter.]* It's kind of funny! Like, what? How could I think that? *[This is an initial marker indicating that the pro-symptom schema may have been successfully disconfirmed, depotentiated, and dissolved by the transformation sequence.]*

Therapist: Do you mean, how could you think that simply saying what you know, or mentioning some good idea that you've had, would make you seem arrogant, insensitive and dominating like Dad and be hated for it?

Richard: *[Laughing again.]* Yeah!

Afterwards, the therapist and Richard wrote a new card together, which Richard was told to review daily:

> All along it's been so clear that if I confidently say what I know, I will always come across as arrogant, insensitive, and dominating like Dad, and be hated for it. And it's so weird, looking around the room and seeing that it doesn't come across like that.

The purpose of the card was to provide additional juxtaposition experiences between the old schema and the new knowledge. While the original transformation sequence might have been enough to eliminate the old schema, the schema might also have been stored in the context of many different situations and contained in several memory systems. In such a situation, further juxtapositions would have helped deal with it.

In a follow-up meeting, Richard reported having lost the feelings of self-doubt, and that speaking up no longer felt like it was any big deal. To verify that the old schema really had lost its power, the therapist tried deliberately provoking his old fears again:

> Dropping his voice to a quieter tone, the therapist added, "But tell me, when you have something to say and just say it, what about the danger of coming across as a know-it-all, like Dad, and being hated for that? What about your fear of that and how urgent it is to protect yourself from that?" […]
>
> Richard took in the question, gazed at the therapist in silence for a few seconds, and then replied, "Well, I don't know what to tell you. All I can say is, that doesn't trouble me any more. And hearing you say it, it seems a little strange that it ever did—like, what was my problem?"

Applied reconsolidation: the schema update process in general

Now that we have looked at a specific example, we can look at a more general version of the process.

Accessing sequence

In Coherence Therapy, the *accessing sequence* is the preliminary phase of making both a person's implicit schema and some disconfirming knowledge accessible, so that they can be used in the juxtaposition process:

1. *Symptom identification.* Establishing which specific symptoms the person regards as problematic, and when and where they manifest. In Richard's case, the general symptom was a lack of confidence, which specifically manifested as negative self-talk in meetings.
2. *Retrieval of target learning.* Bringing into explicit awareness the purpose behind the symptoms. This can then be used to guide the search for disconfirming knowledge, as well as accessing the original schema in order to reconsolidate it. In Richard's case, the purpose was to avoid expressing confidence in a way that would make people hate him.
3. *Identification of disconfirming knowledge.* Identifying some past or present experience which directly contradicts the original learning. This knowledge does not necessarily need to feel "better" or "more positive" than the old one, just as long as it is mutually exclusive with the old one. In Richard's case, the disconfirming knowledge was the experience of his co-worker confidently proposing a solution and being well-received.

Erasure sequence

Once both the target schema and the disconfirming knowledge are known, the erasure steps can be applied to update the learning:

1. *Reactivation of the target schema.* Tapping into the felt truth of the original learning, experiencing it as vividly as possible.
2. *Activation of disconfirming knowledge, mismatching the target schema.* Activating, at the same time, the contradictory belief and having the experience of simultaneously believing in two different things which cannot both be true.
3. *Repetitions of the target-disconfirmation pairing.*

Something that the authors emphasize is that when the target schema is activated, there should be no attempt to explicitly argue against it or disprove it, as this risks pushing it down. Rather, the belief update happens when one experiences their old schema as vividly true, while *also* experiencing an entirely opposite belief as vividly true. It is the juxtaposition of believing X and not-X at the same time, which triggers an inbuilt contradiction-detection mechanism in the brain and forces a restructuring of one's belief system to eliminate the inconsistency.

The book notes that this distinguishes Coherence Therapy from approaches such as Cognitive Behavioral Therapy, which is premised on treating some beliefs as intrinsically irrational and then seeking to disprove them. While UtEB does not go further into the comparison, I note that this is a common complaint that I have heard of CBT: that by defaulting to negative emotions being caused by belief distortions, CBT risks belittling those negative emotions which are actually produced by *correct* evaluations of the world.

I would personally add that not only does treating all of your beliefs - including emotional ones - as provisionally valid seem to be a requirement for actually updating them, this approach is also good rationality practice. After all, you can only seek evidence to test a theory, not confirm it.

If you notice different parts of your mind having conflicting models of how the world works, the correct epistemic stance should be that you are trying to figure out which one is true - not privileging one of them as "more rational" and trying to disprove the other. Otherwise it will be unavoidable that your preconception will cause you to dismiss as false beliefs which are actually true. (Of course, you can still reasonably *anticipate* the belief update going a particular way - but you need to take seriously at least the *possibility* that you will be shown wrong.)

This can actually be a relief. Trying to stack the deck towards receiving favorable evidence would just also sabotage the brain's belief update process. So you might as well give up trying to do so, relax, and just let the evidence come in.

I speculate that this limitation might also be in place in part to help avoid the error where you decide which one of two models is more correct, and then discard the other model entirely. Simultaneously running two contradictory schemas at the same time allows both of them to be properly evaluated and merged rather than one of them being thrown away outright. I suspect that in Richard's case, the resulting process didn't cause him to entirely discard the notion that some behaviors will make him hated like his dad was - it just removed the overgeneralization which had been produced by having too little training data as the basis of the schema.

Of course, this means that there does need to be some contradictory information available which *could* be used to disprove the original schema. One might have a schema for which no disconfirmation is available because it is correct, or a schema which might or might not be correct but which is making things worse and cannot easily be disconfirmed. UtEB mentions the example of a man, "Tómas", who had a desire to be understood and validated by someone important in his life. Tómas remarked that a professional therapist who was being paid for his empathy could never fulfill that role. The update contradicting the schema that nobody in his life really understood him, would have to come from someone actually in his life.

Another issue that may pop up with the erasure sequence is that there is another schema which predicts that, for whatever reason, running this transformation may produce adverse effects. In that case, one needs to address the objecting schema first, essentially be carrying out the entire process on it before returning to the original steps. (This is similar to the phenomenon in e.g. Internal Family Systems, where objecting parts may show up and have their concerns addressed before work on the original part can proceed.)

Verification step

Finally, after the erasure sequence has been run, one seeks to verify that lasting change has indeed happened and that the target schema has been transformed. UtEB offers the following behavioral markers as signs that a learning which has previously generated emotional responses has in fact been erased:

- "A specific emotional reaction abruptly can no longer by reactivated by cues and triggers that formerly did so or by other stressful situations."

- “Symptoms of behavior, emotion, somatics, or thought that were expressions of that emotional reaction also disappear permanently.”
- “Non-recurrence of the emotional reaction and symptoms continues effortlessly and without counteractive or preventive measures of any kind.”

The authors interpret current neuroscience to say that only memory reconsolidation can produce these kinds of markers. They cannot be produced by counteractive or competitive processes, such as trying to learn an opposite habit to replace a neurotic behavior. Counteractive processes are generally fragile and susceptible to relapse. When these markers are observed in clinical work, UtEB argues that one may infer that reconsolidation has led to the original learning being replaced.

Further examples

For additional examples of the schema update process, I recommend reading the book, which contains several more case studies of issues which were dealt with using this approach. Here’s a brief summary of the most detailed ones (note that some of these examples are actually more detailed and include additional complications, such as more than one symptom-producing schema; I have only summarized the most prominent ones to give a taste of them):

- “Charlotte”. *Issue*: obsessive attachment to a former lover. *Schema:* “It would be much better if I was merged with my lover”. *Contradictory knowledge:* The harm caused by not having boundaries.
- “Ted”. *Issue:* an inability to hold a steady job and a general lack of success in life. *Schema: “*If my life is a mess, my father will be forced to admit how badly he screwed up as a parent.” *Contradictory knowledge:* Realizing that Ted’s father would never admit failure, no matter what.
- “Brenda”. *Issue:* stage fright when having a leading role in an upcoming play. *Schema:* being on the stage in front of an audience means being unable to get off, causing helplessness similar to when Brenda was in a car with her alcoholic father and couldn’t get off. *Contradictory knowledge:* re-imagining the scene and the way how Brenda *could* actually have gotten out of the car.
- “Travis”. *Issue:* inability to experience intimate emotional closeness in relationships. *Schema:* “Nobody will pay attention to how I feel or give me understanding for how I’m hurting. I don’t matter, and I’m all on my own.” *Contradictory knowledge:* the therapist’s empathic presence and listening.
- “Regina”. *Issue:* strong anxiety and panic during/after interacting with other people. *Schema:* “I’m acceptable and lovable only if I do everything *perfectly*.” *Contradictory knowledge:* Regina’s Uncle Theo loves her regardless of what imperfections she might have.
- “Carol”. *Issue:* wanting to avoid sex with her husband despite feeling emotionally close to him. *Schema:* engaging in any sexuality means being overtly sexual and harming Carol’s daughter, in the way that Carol was harmed by her mother’s overt sexuality. *Contradictory knowledge:* once the schema was made conscious, it activated the brain’s spontaneous mismatch detection mechanisms and started to feel silly.

As the last item suggests, sometimes just making a schema explicit is enough to start to dismantle it. The authors suggest that the brain has a built-in detection system which compares any consciously experienced beliefs for inconsistencies with other things that a person knows, and can spontaneously create juxtaposition experiences by bringing up such inconsistent information. They suggest that therapies which are

based on digging up previously unconscious material, but which do not have an explicit juxtaposition step, work to the extent that the uncovered material happens to trigger this spontaneous mismatch detection. (We already saw this happening with Richard - once his underlying schema had been made conscious, he was startled to later notice what seemed like a contradiction.)

One may note the connection to the model in *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. If the object is an explicit belief, then any subsystem which is paying attention to that object may presumably detect inconsistencies with that subsystem's own models.

Besides these case studies from Coherence Therapy, the authors also analyze published case studies from Accelerated Experiental Dynamic Psychotherapy, Emotion-Focused Therapy, Eye-Movement Desensitization and Reprocessing, and Interpersonal Neurobiology. They try to show how these cases also carried out a juxtaposition process, even if the theoretical frameworks of those therapies did not explicitly realize it. It is the claim of the authors that any therapy which causes lasting emotional change does it through reconsolidation. Finally, the book contains four essays from other therapists (using Coherence Therapy and EMDR), who analyze some of their own case studies.

Evaluating the book's plausibility

Now that we have looked at the book's claims, let's look at whether we should believe in them.

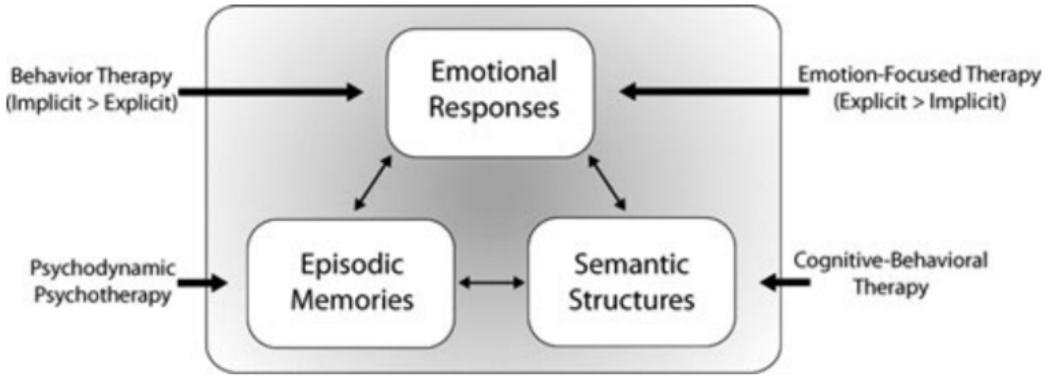
It is unclear to me how reliable the neuroscience results are; the authors cite a number of studies, but each individual claim only references a relatively small number of them.

On a brief look, I could not find any reviews or papers that would have directly made a critical assessment of the book's model. However, I found something that might be even better.

Behavioral and Brain Sciences is a respected journal covering subject areas across the cognitive sciences. BBS publishes "target articles" which present some kind of a thesis or review about a particular topic, together with tens of brief commentaries which respond to the target article, and a final response by the target article's authors to the commentaries.

In 2015, four prestigious (with a total of 500 published research articles between them) psychologists published a BBS target article, *Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science* (Lane et al. 2015). While the exact model that they outline has a number of differences from the UtEB model, the core idea is the same: that therapeutic change from a wide variety of therapeutic approaches, "including behavioral therapy, cognitive-behavioral therapy, emotion-focused therapy, and psychodynamic psychotherapy, results from the updating of prior emotional memories through a process of reconsolidation that incorporates new emotional experiences."

One interesting difference was that Lane et al. describe emotional schemas somewhat differently. In their model, the schemas form memory structures with three mutually integrated components: emotional responses, episodic/autobiographical memories, and semantic structures (e.g. abstract beliefs which generalize over the various incidents, such as the claim that “people are untrustworthy”). Any of these components can be used as an entry point to the memory structure, and can potentially update the other components through reconsolidation. They hypothesize that different forms of therapy work by accessing different types of components: e.g. behavior therapy and emotion-focused therapy access emotional responses, conventional psychoanalysis uses access to biographical memories, and cognitive behavioral therapy accesses semantic structures.



I read the target article, all the commentaries, and the responses. Given the similarities between Lane et al.’s model and the UtEB model, I think we can consider the responses to Lane et al. to generally offer a useful evaluation of the UtEB model as well.

One significant difference which needs to be noted is that Lane et al.’s model of memory reconsolidation does not mention the requirement for a prediction mismatch before reconsolidation can happen. This was remarked on in the response from UtEB’s authors. In their counter-response, Lane et al. noted UtEB’s model to be highly compatible with theirs, and remarked that further research is needed to nail down the conditions which make reconsolidation the most effective.

The other responses to Lane et al. were mostly from psychologists, psychiatrists, and neuroscientists, but also included the occasional economist, philosopher, philologist and folklorist. Several of the responses were generally positive and mostly wanted to contribute additional details or point out future research directions.

However, there were also a number of skeptical responses. A common theme which emerged from several concerned the limitations of the current neuroscience research on memory reconsolidation. In particular, most of the studies so far have been carried out on rats, and specifically testing the elimination of a fear response to electric shocks. As one of the responses points out, “neither the stimuli nor the subjects are generalizable to the kind of rich autobiographical memories involved in therapy.” A number also raised the question whether all therapeutic change really involves reconsolidation, as opposed to some related mechanism, such as creating new memory structures which compete with the original as opposed to replacing it.

My non-expert reading is that the critical responses are right in that a gap remains between the clinical and behavioral findings on the other hand, and the neuroscience

findings on the other. There are various patterns which can be derived from psychological research and clinical therapy experience, and a small number of neuroscience findings which establish the existence of something that *could* explain those patterns. However, the neuroscience findings have only been established in a rather narrow and limited context; the connection between them and the higher-level patterns is a plausible link, but it remains speculative nonetheless.

Personally I consider the book's model tentatively promising, because it seems to explain many observations which I had independently arrived at *before* reading it. For example, I had noticed an interesting thing with anxieties, where I let e.g. a sensation of social anxiety stay active in my mind, neither accepting it as truth *nor* pushing it away while I went to do social things. This would then cause the anxiety to update, making me feel less anxious if it was indeed the case that the social interaction was harmless. This fits nicely together with the framework of an activated memory structure becoming open to reconsolidation and then being updated by a prediction mismatch (the situation not being as bad as expected).

Likewise, in my post Integrating disagreeing subagents, I reviewed a variety of rationality and therapeutic techniques, and suggested that they mostly worked either by merging or combining two existing models that a person's brain already had, or augmenting the existing models by collecting additional information.

In particular, I considered an example from cognitive behavioral therapy, where a man named Walter was feeling like he was impossible to be in a relationship with after he had broken up with his boyfriend. At the same time, he did not think that someone *else* breaking up with their partner was an indication of them being impossible to be in a relationship with. He and his therapist role-played an interaction where the therapist pretended to be a friend who had recently broken up, and Walter explained why this did not make the friend a relationship failure. In the process of doing so, Walter suddenly realized that he wasn't a failure, either.

I commented:

> 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.

This seems like a straightforward instance of belief juxtaposition, and one where I ended up independently deriving something like UtEB's memory reconsolidation model: I too noted that the relevant belief structures need to be simultaneously activated in the right way to allow for the brain to revise one of them after noticing the contradiction. In general, UtEB's model of how things work rings true in my experience, making me inclined to believe that its description of how therapy works is correct, and that its model of how it is connected to neuroscience might also be.

UtEB and the subagent model

As many readers know, I have been writing a sequence of posts on multi-agent models of mind. In Building up to an Internal Family Systems model, I suggested that the human mind might contain something like subagents which try to ensure that past catastrophes do not repeat. In subagents, coherence, and akrasia in humans, I suggested that behaviors such as procrastination, indecision, and seemingly inconsistent behavior result from different subagents having disagreements over what to do.

As I already mentioned, my post on integrating disagreeing subagents took the model in the direction of interpreting disagreeing subagents as conflicting beliefs or models within a person's brain. Subagents, trauma and rationality further suggested that the appearance of drastically different personalities within a single person might result from unintegrated memory networks, which resist integration due to various traumatic experiences.

This post has discussed UtEB's model of conflicting emotional schemas in a way which further equates "subagents" with beliefs - in this case, the various schemas seem closely related to what e.g. Internal Family Systems calls "parts". In many situations, it is probably fair to say that this is what subagents are.

That said, I think that while this covers a very important subset of subagents, not *everything* which I have been referring to as a subagent falls straightforwardly under the belief-schema model. In subagents and neural Turing machines as well as Against "System 1" and "System 2", I also covered subagents in a more general way, as also including e.g. the kinds of subsystems which carry out object recognition and are used to carry out tasks like arithmetic. This was also the lens through which I looked at subagents in my summary of Consciousness and the Brain. Which kind of view is the most useful, depends on exactly what phenomenon we are trying to understand.

Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More

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

An actual debate about instrumental convergence, in a public space! Major respect to all involved, especially Yoshua Bengio for great facilitation.

For posterity (i.e. having a good historical archive) and further discussion, I've reproduced the conversation here. I'm happy to make edits at the request of anyone in the discussion who is quoted below. I've improved formatting for clarity and fixed some typos. For people who are not researchers in this area who wish to comment, see the public version of this post here. For people who do work on the relevant areas, please sign up in the top right. It will take a day or so to confirm membership.

Original Post

Yann LeCun: "don't fear the Terminator", a short opinion piece by Tony Zador and me that was just published in Scientific American.

"We dramatically overestimate the threat of an accidental AI takeover, because we tend to conflate intelligence with the drive to achieve dominance. [...] But intelligence per se does not generate the drive for domination, any more than horns do."

https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/

Comment Thread #1

Elliot Olds: Yann, the smart people who are very worried about AI seeking power and ensuring its own survival believe it's a big risk because power and survival are instrumental goals for almost any ultimate goal.

If you give a generally intelligent AI the goal to make as much money in the stock market as possible, it will resist being shut down because that would interfere with tis goal. It would try to become more powerful because then it could make money more effectively. This is the natural consequence of giving a smart agent a goal, unless we do something special to counteract this.

You've often written about how we shouldn't be so worried about AI, but I've never seen you address this point directly.

Stuart Russell: It is trivial to construct a toy MDP in which the agent's only reward comes from fetching the coffee. If, in that MDP, there is another "human" who has some probability, however small, of switching the agent off, and if the agent has available a button that switches off that human, the agent will necessarily press that button as part of the optimal solution for fetching the coffee. No hatred, no desire for

power, no built-in emotions, no built-in survival instinct, nothing except the desire to fetch the coffee successfully. This point cannot be addressed because it's a simple mathematical observation.

Comment Thread #2

Yoshua Bengio: Yann, I'd be curious about your response to Stuart Russell's point.

Yann LeCun: You mean, the so-called "instrumental convergence" argument by which "a robot can't fetch you coffee if it's dead. Hence it will develop self-preservation as an instrumental sub-goal."

It might even kill you if you get in the way.

1. Once the robot has brought you coffee, its self-preservation instinct disappears. You can turn it off.

2. One would have to be unbelievably stupid to build open-ended objectives in a super-intelligent (and super-powerful) machine without some safeguard terms in the objective.

3. One would have to be rather incompetent not to have a mechanism by which new terms in the objective could be added to prevent previously-unforeseen bad behavior. For humans, we have education and laws to shape our objective functions and complement the hardwired terms built into us by evolution.

4. The power of even the most super-intelligent machine is limited by physics, and its size and needs make it vulnerable to physical attacks. No need for much intelligence here. A virus is infinitely less intelligent than you, but it can still kill you.

5. A second machine, designed solely to neutralize an evil super-intelligent machine will win every time, if given similar amounts of computing resources (because specialized machines always beat general ones).

Bottom line: there are lots and lots of ways to protect against badly-designed intelligent machines turned evil.

Stuart has called me stupid in the Vanity Fair interview linked below for allegedly not understanding the whole idea of instrumental convergence.

It's not that I don't understand it. I think it would only be relevant in a fantasy world in which people would be smart enough to design super-intelligent machines, yet ridiculously stupid to the point of giving it moronic objectives with no safeguards.

Here is the juicy bit from the article where Stuart calls me stupid:

> Russell took exception to the views of Yann LeCun, who developed the forerunner of the convolutional neural nets used by AlphaGo and is Facebook’s director of A.I. research. LeCun told the BBC that there would be no Ex Machina or Terminator scenarios, because robots would not be built with human drives—hunger, power, reproduction, self-preservation. “Yann LeCun keeps saying that there’s no reason why machines would have any self-preservation instinct,” Russell said. “And it’s simply and mathematically false. I mean, it’s so obvious that a machine will have self-preservation even if you don’t program it in because if you say, ‘Fetch the

> coffee,’ it can’t fetch the coffee if it’s dead. So if you give it any goal whatsoever, it has a reason to preserve its own existence to achieve that goal. And if you threaten it on your way to getting coffee, it’s going to kill you because any risk to the coffee has to be countered. People have explained this to LeCun in very simple terms.”

https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x

Tony Zador: I agree with most of what Yann wrote about Stuart Russell's concern.

Specifically, I think the flaw in Stuart's argument is the assertion that "switching off the human is the optimal solution"---who says that's an optimal solution?

I guess if you posit an omnipotent robot, destroying humanity might be a possible solution. But if the robot is not omnipotent, then killing humans comes at considerable risk, ie that they will retaliate. Or humans might build special "protector robots" whose value function is solely focused on preventing the killing of humans by other robots. Presumably these robots would be at least as well armed as the coffee robots. So this really increases the risk to the coffee robots of pursuing the genocide strategy.

And if the robot is omnipotent, then there are an infinite number of alternative strategies to ensure survival (like putting up an impenetrable forcefield around the off switch) that work just as well.

So i would say that killing all humans is not only not likely to be an optimal strategy under most scenarios, the set of scenarios under which it is optimal is probably close to a set of measure 0.

Stuart Russell: Thanks for clearing that up - so 2+2 is not equal to 4, because if the 2 were a 3, the answer wouldn't be 4? I simply pointed out that in the MDP as I defined it, switching off the human is the optimal solution, despite the fact that we didn't put in any emotions of power, domination, hate, testosterone, etc etc. And your solution seems, well, frankly terrifying, although I suppose the NRA would approve. Your last suggestion, that the robot could prevent anyone from ever switching it off, is also one of the things we are trying to avoid. The point is that the behaviors we are concerned about have nothing to do with putting in emotions of survival, power, domination, etc. So arguing that there's no need to put those emotions in is completely missing the point.

Yann LeCun: Not clear whether you are referring to my comment or Tony's.

The point is that behaviors you are concerned about are easily avoidable by simple terms in the objective. In the unlikely event that these safeguards somehow fail, my partial list of escalating solutions (which you seem to find terrifying) is there to prevent a catastrophe. So arguing that emotions of survival etc will inevitably lead to dangerous behavior is completely missing the point.

It's a bit like saying that building cars without brakes will lead to fatalities.

Yes, but why would we be so stupid as to not include brakes?

That said, instrumental subgoals are much weaker drives of behavior than hardwired objectives. Else, how could one explain the lack of domination behavior in non-social animals, such as orangutans.

Francesca Rossi: @Yann Indeed it would be odd to design an AI system with a specific goal, like fetching coffee, and capabilities that include killing humans or disallowing being turned off, without equipping it also with guidelines and priorities to constrain its freedom, so it can understand for example that fetching coffee is not so important that it is worth killing a human being to do it. Value alignment is fundamental to achieve this. Why would we build machines that are not aligned to our values? Stuart, I agree that it would easy to build a coffee fetching machine that is not aligned to our values, but why would we do this? Of course value alignment is not easy, and still a research challenge, but I would make it part of the picture when we envision future intelligent machines.

Richard Mallah: Francesca, of course Stuart believes we should create value-aligned AI. The point is that there are too many caveats to explicitly add each to an objective function, and there are strong socioeconomic drives for humans to monetize AI prior to getting it sufficiently right, sufficiently safe.

Stuart Russell: "Why would be build machines that are not aligned to our values?" That's what we are doing, all the time. The standard model of AI assumes that the objective is fixed and known (check the textbook!), and we build machines on that basis - whether it's clickthrough maximization in social media content selection or total error minimization in photo labeling (Google Jacky Alciné) or, per Danny Hillis, profit maximization in fossil fuel companies. This is going to become even more untenable as machines become more powerful. There is no hope of "solving the value alignment problem" in the sense of figuring out the right value function offline and putting it into the machine. We need to change the way we do AI.

Yoshua Bengio: All right, we're making some progress towards a healthy debate. Let me try to summarize my understanding of the arguments. Yann LeCun and Tony Zadorr argue that humans would be stupid to put in explicit dominance instincts in our AIs. Stuart Russell responds that it needs not be explicit but dangerous or immoral behavior may simply arise out of imperfect value alignment and instrumental subgoals set by the machine to achieve its official goals. Yann LeCun and Tony Zador respond that we would be stupid not to program the proper 'laws of robotics' to protect humans. Stuart Russell is concerned that value alignment is not a solved problem and may be intractable (i.e. there will always remain a gap, and a sufficiently powerful AI could 'exploit' this gap, just like very powerful corporations currently often act legally but immorally). Yann LeCun and Tony Zador argue that we could also build defensive military robots designed to only kill regular AIs gone rogue by lack of value alignment. Stuart Russell did not explicitly respond to this but I infer from his NRA reference that we could be worse off with these defensive robots because now they have explicit weapons and can also suffer from the value misalignment problem.

Yoshua Bengio: So at the end of the day, it boils down to whether we can handle the value misalignment problem, and I'm afraid that it's not clear we can for sure, but it also seems reasonable to think we will be able to in the future. Maybe part of the problem is that Yann LeCun and Tony Zador are satisfied with a 99.9% probability that we can fix the value alignment problem while Stuart Russell is not satisfied with taking such an existential risk.

Yoshua Bengio: And there is another issue which was not much discussed (although the article does talk about the short-term risks of military uses of AI etc), and which concerns me: humans can easily do stupid things. So even if there are ways to mitigate the possibility of rogue AIs due to value misalignment, how can we guarantee that no single human will act stupidly (more likely, greedily for their own power) and

unleash dangerous AIs in the world? And for this, we don't even need superintelligent AIs, to feel very concerned. The value alignment problem also applies to humans (or companies) who have a lot of power: the misalignment between their interests and the common good can lead to catastrophic outcomes, as we already know (e.g. tragedy of the commons, corruption, companies lying to have you buy their cigarettes or their oil, etc). It just gets worse when more power can be concentrated in the hands of a single person or organization, and AI advances can provide that power.

Francesca Rossi: I am more optimistic than Stuart about the value alignment problem. I think that a suitable combination of symbolic reasoning and various forms of machine learning can help us to both advance AI's capabilities and get closer to solving the value alignment problem.

Tony Zador: @Stuart Russell "Thanks for clearing that up - so 2+2 is not equal to 4, because if the 2 were a 3, the answer wouldn't be 4? "

hmm. not quite what i'm saying.

If we're going for the math analogies, then i would say that a better analogy is:

Find X, Y such that X+Y=4.

The "killer coffee robot" solution is {X=642, Y = -638}. In other words: Yes, it is a solution, but not a particularly natural or likely or good solution.

But we humans are blinded but our own warped perspective. We focus on the solution that involves killing other creatures because that appears to be one of the main solutions that we humans default to. But it is not a particularly common solution in the natural world, nor do i think it's a particularly effective solution in the long run.

Yann LeCun: Humanity has been very familiar with the problem of fixing value misalignments for millenia.

We fix our children's hardwired values by teaching them how to behave.

We fix human value misalignment by laws. Laws create extrinsic terms in our objective functions and cause the appearance of instrumental subgoals ("don't steal") in order to avoid punishment. The desire for social acceptance also creates such instrumental subgoals driving good behavior.

We even fix value misalignment for super-human and super-intelligent entities, such as corporations and governments.

This last one occasionally fails, which is a considerably more immediate existential threat than AI.

Tony Zador: @Yoshua Bengio I agree with much of your summary. I agree value alignment is important, and that it is not a solved problem.

I also agree that new technologies often have unintended and profound consequences. The invention of books has led to a decline in our memories (people used to recite the entire Odyssey). Improvements in food production technology (and other factors) have led to a surprising obesity epidemic. The invention of social media is disrupting our political systems in ways that, to me anyway, have been quite

surprising. So improvements in AI will undoubtedly have profound consequences for society, some of which will be negative.

But in my view, focusing on "killer robots that dominate or step on humans" is a distraction from much more serious issues.

That said, perhaps "killer robots" can be thought of as a metaphor (or metonym) for the set of all scary scenarios that result from this powerful new technology.

Yann LeCun: @Stuart Russell you write "we need to change the way we do AI". The problems you describe have nothing to do with AI per se.

They have to do with designing (not avoiding) explicit instrumental objectives for entities (e.g. corporations) so that their overall behavior works for the common good. This is a problem of law, economics, policies, ethics, and the problem of controlling complex dynamical systems composed of many agents in interaction.

What is required is a mechanism through which objectives can be changed quickly when issues surface. For example, Facebook stopped maximizing clickthroughs several years ago and stopped using the time spent in the app as a criterion about 2 years ago. It put in place measures to limit the dissemination of clickbait, and it favored content shared by friends rather than directly disseminating content from publishers.

We certainly agree that designing good objectives is hard. Humanity has struggled with designing objectives for itself for millennia. So this is not a new problem. If anything, designing objectives for machines, and forcing them to abide by them will be a lot easier than for humans, since we can physically modify their firmware.

There will be mistakes, no doubt, as with any new technology (early jetliners lost wings, early cars didn't have seat belts, roads didn't have speed limits...).

But I disagree that there is a high risk of accidentally building existential threats to humanity.

Existential threats to humanity have to be explicitly designed as such.

Yann LeCun: It will be much, much easier to control the behavior of autonomous AI systems than it has been for humans and human organizations, because we will be able to directly modify their intrinsic objective function.

This is very much unlike humans, whose objective can only be shaped through extrinsic objective functions (through education and laws), that indirectly create instrumental sub-objectives ("be nice, don't steal, don't kill, or you will be punished").

As I have pointed out in several talks in the last several years, autonomous AI systems will need to have a trainable part in their objective, which would allow their handlers to train them to behave properly, without having to directly hack their objective function by programmatic means.

Yoshua Bengio: Yann, these are good points, we indeed have much more control over machines than humans since we can design (and train) their objective function. I actually have some hopes that by using an objective-based mechanism relying on learning (to inculcate values) rather than a set of hard rules (like in much of our legal system), we could achieve more robustness to unforeseen value alignment mishaps.

In fact, I surmise we should do that with human entities too, i.e., penalize companies, e.g. fiscally, when they behave in a way which hurts the common good, even if they are not directly violating an explicit law. This also suggests to me that we should try to avoid that any entity (person, company, AI) have too much power, to avoid such problems. On the other hand, although probably not in the near future, there could be AI systems which surpass human intellectual power in ways that could foil our attempts at setting objective functions which avoid harm to us. It seems hard to me to completely deny that possibility, which thus would beg for more research in (machine-) learning moral values, value alignment, and maybe even in public policies about AI (to minimize the events in which a stupid human brings about AI systems without the proper failsafes) etc.

Yann LeCun: @Yoshua Bengio if we can build "AI systems which surpass human intellectual power in ways that could foil our attempts at setting objective functions", we can also build similarly-powerful AI systems to set those objective functions.

Sort of like the discriminator in GANs....

Yann LeCun: @Yoshua Bengio a couple direct comments on your summary:

- designing objectives for super-human entities is not a new problem. Human societies have been doing this through laws (concerning corporations and governments) for millennia.
- the defensive AI systems designed to protect against rogue AI systems are not akin to the military, they are akin to the police, to law enforcement. Their "jurisdiction" would be strictly AI systems, not humans.

But until we have a hint of a beginning of a design, with some visible path towards autonomous AI systems with non-trivial intelligence, we are arguing about the sex of angels.

Yuri Barzov: Aren't we overestimating the ability of imperfect humans to build a perfect machine? If it will be much more powerful than humans its imperfections will be also magnified. Cute human kids grow up into criminals if they get spoiled by reinforcement i.e. addiction to rewards. We use reinforcement and backpropagation (kind of reinforcement) in modern golden standard AI systems. Do we know enough about humans to be able to build a fault-proof human friendly super intelligent machine?

Yoshua Bengio: @Yann LeCun, about discriminators in GANs, and critics in Actor-Critic RL, one thing we know is that they tend to be biased. That is why the critic in Actor-Critic is not used as an objective function but instead as a baseline to reduce the variance. Similarly, optimizing the generator wrt a fixed discriminator does not work (you would converge to a single mode - unless you balance that with entropy maximization). Anyways, just to say, there is much more research to do, lots of unknown unknowns about learning moral objective functions for AIs. I'm not afraid of research challenges, but I can understand that some people would be concerned about the safety of gradually more powerful AIs with misaligned objectives. I actually like the way that Stuart Russell is attacking this problem by thinking about it not just in terms of an objective function but also about uncertainty: the AI should avoid actions which might hurt us (according to a self-estimate of the uncertain consequences of actions), and stay the conservative course with high confidence of accomplishing the mission while not creating collateral damage. I think that what you and I are trying to say is that all this is quite different from the terminator scenarios

which some people in the media are brandishing. I also agree with you that there are lots of unknown unknowns about the strengths and weaknesses of future AIs, but I think that it is not too early to start thinking about these issues.

Yoshua Bengio: @Yuri Barzov the answer to your question: no. But we don't know that it is not feasible either, and we have reasons to believe that (a) it is not for tomorrow such machines will exist and (b) we have intellectual tools which may lead to solutions. Or maybe not!

Stuart Russell: Yann's comment "Facebook stopped maximizing clickthroughs several years ago and stopped using the time spent in the app as a criterion about 2 years ago" makes my point for me. Why did they stop doing it? Because it was the wrong objective function. Yann says we'd have to be "extremely stupid" to put the wrong objective into a super-powerful machine. Facebook's platform is not super-smart but it is super-powerful, because it connects with billions of people for hours every day. And yet they put the wrong objective function into it. QED. Fortunately they were able to reset it, but unfortunately one has to assume it's still optimizing a fixed objective. And the fact that it's operating within a large corporation that's designed to maximize another fixed objective - profit - means we cannot switch it off.

Stuart Russell: Regarding "externalities" - when talking about externalities, economists are making essentially the same point I'm making: externalities are the things not stated in the given objective function that get damaged when the system optimizes that objective function. In the case of the atmosphere, it's relatively easy to measure the amount of pollution and charge for it via taxes or fines, so correcting the problem is possible (unless the offender is too powerful). In the case of manipulation of human preferences and information states, it's very hard to assess costs and impose taxes or fines. The theory of uncertain objectives suggests instead that systems be designed to be "minimally invasive", i.e., don't mess with parts of the world state whose value is unclear. In particular, as a general rule it's probably best to avoid using fixed-objective reinforcement learning in human-facing systems, because the reinforcement learner will learn how to manipulate the human to maximize its objective.

Stuart Russell: @Yann LeCun Let's talk about climate change for a change. Many argue that it's an existential or near-existential threat to humanity. Was it "explicitly designed" as such? We created the corporation, which is a fixed-objective maximizer. The purpose was not to create an existential risk to humanity. Fossil-fuel corporations became super-powerful and, in certain relevant senses, super-intelligent: they anticipated and began planning for global warming five decades ago, executing a campaign that outwitted the rest of the human race. They didn't win the academic argument but they won in the real world, and the human race lost. I just attended an NAS meeting on climate control systems, where the consensus was that it was too dangerous to develop, say, solar radiation management systems - not because they might produce unexpected disastrous effects but because the fossil fuel corporations would use their existence as a further form of leverage in their so-far successful campaign to keep burning more carbon.

Stuart Russell: @Yann LeCun This seems to be a very weak argument. The objection raised by Omohundro and others who discuss instrumental goals is aimed at any system that operates by optimizing a fixed, known objective; which covers pretty much all present-day AI systems. So the issue is: what happens if we keep to that general plan - let's call it the "standard model" - and improve the capabilities for the system to achieve the objective? We don't need to know today *how* a future system

achieves objectives more successfully, to see that it would be problematic. So the proposal is, don't build systems according to the standard model.

Yann LeCun: @Stuart Russell the problem is that essentially no AI system today is autonomous.

They are all trained *in advance* to optimize an objective, and subsequently execute the task with no regards to the objective, hence with no way to spontaneously deviate from the original behavior.

As of today, as far as I can tell, we do *not* have a good design for an autonomous machine, driven by an objective, capable of coming up with new strategies to optimize this objective in the real world.

We have plenty of those in games and simple simulation. But the learning paradigms are way too inefficient to be practical in the real world.

Yuri Barzov: @Yoshua Bengio yes. If we frame the problem correctly we will be able to resolve it. AI puts natural intelligence into focus like a magnifying mirror

Yann LeCun: @Stuart Russell in pretty much everything that society does (business, government, of whatever) behaviors are shaped through incentives, penalties via contracts, regulations and laws (let's call them collectively the objective function), which are proxies for the metric that needs to be optimized.

Because societies are complex systems, because humans are complex agents, and because conditions evolve, it is a requirement that the objective function be modifiable to correct unforeseen negative effects, loopholes, inefficiencies, etc.

The Facebook story is unremarkable in that respect: when bad side effects emerge, measures are taken to correct them. Often, these measures eliminate bad actors by directly changing their economic incentive (e.g. removing the economic incentive for clickbaits).

Perhaps we agree on the following:

(0) not all consequences of a fixed set of incentives can be predicted.

(1) because of that, objectives functions must be updatable.

(2) they must be updated to correct bad effect whenever they emerge.

(3) there should be an easy way to train minor aspects of objective functions through simple interaction (similar to the process of educating children), as opposed to programmatic means.

Perhaps where we disagree is the risk of inadvertently producing systems with badly-designed and (somehow) un-modifiable objectives that would be powerful enough to constitute existential threats.

Yoshua Bengio: @Yann LeCun this is true, but one aspect which concerns me (and others) is the gradual increase in power of some agents (now mostly large companies and some governments, potentially some AI systems in the future). When it was just weak humans the cost of mistakes or value misalignment (improper laws, misaligned objective function) was always very limited and local. As we build more and more

powerful and intelligent tools and organizations, (1) it becomes easier to cheat for 'smarter' agents (exploit the misalignment) and (2) the cost of these misalignments becomes greater, potentially threatening the whole of society. This then does not leave much time and warning to react to value misalignment.

What Comes After Epistemic Spot Checks?

When I read a non-fiction book, I want to know if it's correct before I commit anything it says to memory. But if I already knew the truth status of all of its claims, I wouldn't need to read it. Epistemic Spot Checks are an attempt to square that circle by sampling a book's claims and determining their truth value, with the assumption that the sample is representative of the whole.

Some claims are easier to check than others. On one end are simple facts, e.g., "Emperor Valerian spent his final years as a Persian prisoner". This was easy and quick to verify: googling "emperor valerian" was quite sufficient. "Roman ship sizes weren't exceeded until the 15th century" looks similar, but it wasn't. If you google the claim itself, it will look confirmed (evidence: me and 3 other people in the forecasting tournament did this). At the last second while researching this, I decided to check the size of Chinese ships, which surpassed Roman ships sooner than Europe did, by about a century.

On first blush this looks like a success story, but it's not. I was only able to catch the mistake because I had a bunch of background knowledge about the state of the world. If I didn't already know mid-millenium China was better than Europe at almost everything (and I remember a time when I didn't), I could easily have drawn the wrong conclusion about that claim. And following a procedure that would catch issues like this every time would take much more time than ESCs currently get.

Then there's terminally vague questions, like "Did early modern Europe have more emphasis on rationality and less superstition than other parts of the world?" (As claimed by The Unbound Prometheus). It would be optimistic to say that question requires several books to answer, but even if that were true, each of them would need at least an ESC themselves to see if they're trustworthy, which might involve checking other claims requiring several books to verify… pretty soon it's a master's thesis.

But I can't get a master's degree in everything interesting or relevant to my life. And that brings up another point: credentialism. A lot of ESC revolves around "Have people who have officially been Deemed Credible sanctioned this fact?" rather than "Have I seen evidence that I, personally, judge to be compelling?"

The Fate of Rome (Kyle Harper) and *The Fall of Rome* (Bryan Ward-Perkins) are both about the collapse of the western Roman empire. They both did almost flawlessly on their respective epistemic spot checks. And yet, they attribute the fall of Rome to very different causes, and devote almost no time to the others' explanation. If you drew a venn diagram of the data they discuss, the circles would be almost but not quite entirely distinct. The word "plague" appears 651 times in *Fate* and 6 times in *Fall*, who introduces the topic mostly to dismiss the idea that it was causally related to the fall-which is how *Fate* treats all those border adjustments happening from 300 AD on. *Fate* is very into discussing climate, but *Fall* uses that space to talk about pottery.

This is why I called the process epistemic spot checking, not truth-checking. Determining if a book is true requires not only determining if each individual claim is true, but what other explanations exist and what has been left out. Depending on the specifics, ESC as I do them now are perhaps half the work of reading the subsection of

the book I verify. Really thoroughly checking a single paragraph in a published paper took me 10-20 hours. And even if I succeed at the ESC, all I have is a thumbs up/thumbs down on the book.

Around the same time I was doing ESCs on *The Fall of Rome* and *The Fate of Rome* (the ESCs were published far apart to get maximum publicity for the Amplification experiment, but I read and performed the ESCs very close together), I was commissioned to do a shallow review on the question of "How to get consistency in predictions or evaluations of questions?" I got excellent feedback from the person who commissioned it, but I felt like it said a lot more about the state of a field of literature than the state of the world, because I had to take authors' words for their findings. It had all the problems ESCs were designed to prevent.

I'm in the early stages of trying to form something better: something that incorporates the depth of epistemic spot checks with the breadth of field reviews. It's designed to bootstrap from knowing nothing in a field to being grounded and informed and having a firm base on which to evaluate new information. This is hard and I expect it to take a while.

Turning air into bread

This is a linkpost for https://rootsofprogress.org/turning-air-into-bread

Originally posted on The Roots of Progress, August 12, 2017

I recently finished *The Alchemy of Air*, by Thomas Hager. It's the story of the Haber-Bosch process, the lives of the men who created it, and its consequences for world agriculture and for Germany during the World Wars.

What is the Haber-Bosch process? It's what keeps billions of people in the modern world from starving to death. In Hager's phrase: it turns air into bread.

Some background. Plants, like all living organisms, need to take in nutrients for metabolism. For animals, the macronutrients needed are large, complex molecules: proteins, carbohydrates, fats. But for plants they are elements: nitrogen, phosphorus and potassium (NPK). Nitrogen is needed in the largest quantities.

Nitrogen is all around us: it constitutes about four-fifths of the atmosphere. But plants can't use atmospheric nitrogen. Nitrogen gas, N_2, consists of two atoms held together by a triple covalent bond. The strength of this bond renders nitrogen mostly inert: it doesn't react with much. To use it in chemical processes, plants need other nitrogen-containing molecules. These substances are known as “fixed” nitrogen; the process of turning nitrogen gas into usable form is called fixation.

In nature, nitrogen fixation is performed by bacteria. Some of these bacteria live in the soil; some live in a symbiotic relationship on the roots of certain plants, such as peas and other legumes.

Nitrogen availability is one of the top factors in plant growth and therefore in agriculture. The more fixed nitrogen is in the soil, the more crops can grow. Unfortunately, when you farm a plot of land, natural processes don't replace the nitrogen as fast as it is depleted.

Pre-industrial farmers had no chemistry or advanced biology to guide them, but they knew that soil would lose its fertility over the years, and they had learned a few tricks. One was fertilization with natural substances, particularly animal waste, which contains nitrogen. Another was crop rotation: planting peas, for instance, would replace some of the nitrogen in the soil, thanks to those nitrogen-fixing bacteria on their roots.

But these techniques could only go so far. As the world population increased in the 19th century, more and more farmland was needed. Famine was staved off, for a time, by the opening of the prairies of the New World, but those resources were finite. The world needed fertilizer.

An island off the coast of Peru where it almost never rains had accumulated untold centuries of—don't laugh—seagull droppings, some of the world's best known natural fertilizer. An industry was made out of mining guano on these islands, where it was piled several stories high, and shipping it all over the world. When that ran out after a couple decades, attention turned inland to the Atacama Desert, where, with no rainfall

and no life, unusual minerals grew in crystals on the rocks. The crystals included *salitre*, or Chilean saltpeter, a nitrogen salt that could be made into fertilizer.

It could be made into something else important, too: gunpowder. It turns out that nitrogen is a crucial component not only of fertilizer, but also of explosives. Needing it both to feed and to arm their people, every country considered saltpeter a strategic commodity. Peru, Chile and Bolivia went to war over the saltpeter resources of the Atacama in the late 1800s (Bolivia, at the time, had a small strip of land in the desert, running to the ocean; it lost that strip in the war and has remained landlocked ever since).

By the end of the 19th century, as population continued to soar, it was clear that the Chilean saltpeter would run out within decades, just as the guano had. Sir William Crookes, head of the British Academy of Sciences, warned that the world was heading for mass famine, a true Malthusian catastrophe, unless we discovered a way to synthesize fertilizer. And he called on the chemists of the world to do it.

Nearby, in Germany, other scientists were thinking the same thing. Germany was highly dependent on salt shipped halfway around the world from Chile. But Germany did not have the world's best navy. If—God forbid—Germany were ever to be at war with England (!), they would quickly blockade Germany and deprive it of nitrogen. Germany would have no food and no bombs—not a good look, in wartime.

The prospect of synthesizing fixed nitrogen was tantalizing. After all, the nitrogen itself is abundant in the atmosphere. A product such as ammonia, NH_3, could be made from that and hydrogen, which of course is present in water. All you need is a way to put them together in the right combination.

The problem, again, is that triple covalent bond. Owing to the strength of that bond, it takes very high temperatures to rip N_2 apart. More troublesome is that ammonia is by comparison a weak molecule. So at temperatures high enough to separate the nitrogen atoms, the ammonia basically burns up.

Fritz Haber was the chemist who solved the fundamental problem. He found that increasing the pressure of the gases allowed him to decrease the temperature. At very high pressures, he could start to get an appreciable amount of ammonia. By introducing the right catalyst, he could increase the production to levels that were within reach of a viable industrial process.

Carl Bosch was the industrialist at the German chemical company BASF who led the team that figured out how to turn this into a profitable process, at scale. The challenges were enormous. To start with, the pressures required by the process were immense, around 200 atmospheres. The required temperatures, too, were very high. No one had ever done industrial chemistry in that regime before, and Bosch's team had to invent almost everything from scratch, pioneering an entirely new subfield of high-pressure industrial chemistry. Their furnaces kept exploding—not only from the pressure itself, but because hydrogen was eating away at the steel walls of the container, as it forced into them. No material was strong enough and inexpensive enough to serve as the container wall. Finally Bosch came up with an ingenious system in which the furnaces had an inner lining of material to protect the steel, which would be replaced on a regular basis.

A further challenge was the catalyst: Haber had used osmium, an extremely rare metal. BASF bought up the entire world's supply, but it wasn't enough to produce the quantities they needed. They experimented with thousands of other materials, finally settling on a catalyst with an iron base combined with other elements.

This is the Haber-Bosch process: it turns pure nitrogen and hydrogen gas into ammonia. The nitrogen can be isolated from the atmosphere (by cooling air until it condenses into liquid, then carefully increasing the temperature: different substances boil at different temperatures, so this process separates them). Hydrogen can be produced from water by electrolysis, or, these days, found in natural gas deposits. The output of the process, ammonia, is the precursor of many important products, including fertilizers and explosives.

The new BASF plant that Bosch built began turning out tons of ammonia a day. It beat out all competing processes (including one that used electric arcs through the air), and provided the world with fertilizer—cheaper and of more consistent quality than could be obtained from the salts of Chile, which were abandoned before they ran out.

Haber-Bosch fed the world—but it also prolonged World War I, and later helped fuel the rise of Hitler.

The Alchemy of Air is as much about the lives of Haber and Bosch, and what happened after their process became a reality, as it is about the science and technology of the process itself. Even though the technology was my main interest this time, I found the history captivating.

Haber was a Jew, at a time when Jews were second-class citizens in Germany. Rather than denouncing the society he lived in, this seemed to cause Haber to seek its approval. After his scientific achievement with ammonia, he got a high-status job at the Kaiser Wilhelm Institutes in Berlin, and sought to be an adviser to the Kaiser himself. Jews were barred from military service, but Haber was able to become a science adviser to the military—even pioneering the use of poison gas in WW1, a role that left him with a reputation as a war criminal.

Haber believed that if Jews showed what good, patriotic German citizens they could be, they could eventually be accepted as equals. Decades later, when the Nazis came to power and began “cleansing” Jews first out of the German government, then out of all of society, Haber saw his dream of acceptance fall completely to pieces. He died, shortly before WW2, in great distress.

Bosch, on the other hand, held liberal political views and was against the Nazis. He even tried to speak out against them, and in a personal meeting with Hilter made a futile argument for freedom of inquiry and better treatment of the Jews. But at the same time he made deals with the Nazis to secure funding for his chemical company —by then he was the head, not only of BASF, but of a broader industry association called IG Farben. He was building a massive chemical plant in the heart of Germany, at Leuna, to produce not only ammonia but also what he saw as his *magnum opus:* synthetic gasoline, made from coal. In the end Farben became virtually a state company and provided much of the material Germany needed for WW2, including ammonia, gasoline, and rubber.

Bosch died shortly after the war began. On his deathbed, he predicted that the war would be a disaster for Germany. It would go well at first, he said, and Germany would

occupy France and maybe even Britain. But then Hitler would make the fatal mistake of invading Russia. In the end, the skies would darken with Allied planes, and much of Germany would be destroyed. It happened as he predicted, and Bosch's beloved Leuna was a major target, ultimately crippled by wave after wave of Allied bombing raids.

Synthetic ammonia is one of the most important industrial products of the modern world, and so Haber-Bosch is one of the most important industrial processes. Around 1% of the *total* energy of the economy is devoted to it, and Hager estimates that half the nitrogen atoms in your body came from it. It's a crucial part of the story of industrial agriculture, and so a crucial part of the story of how we became smart, rich and free.

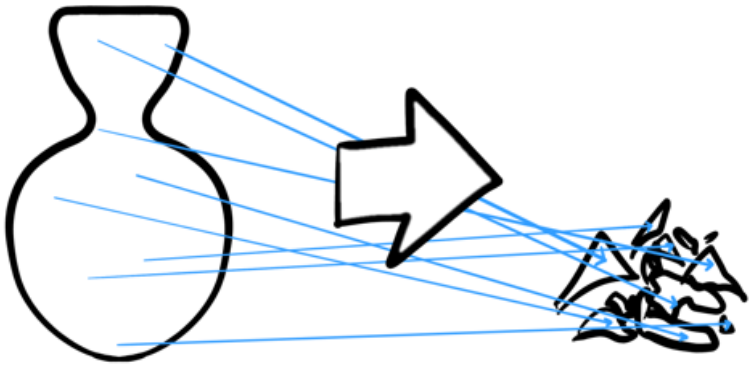
The Alchemy of Air: A Jewish Genius, a Doomed Tycoon, and the Scientific Discovery That Fed the World but Fueled the Rise of Hitler

World State is the Wrong Abstraction for Impact

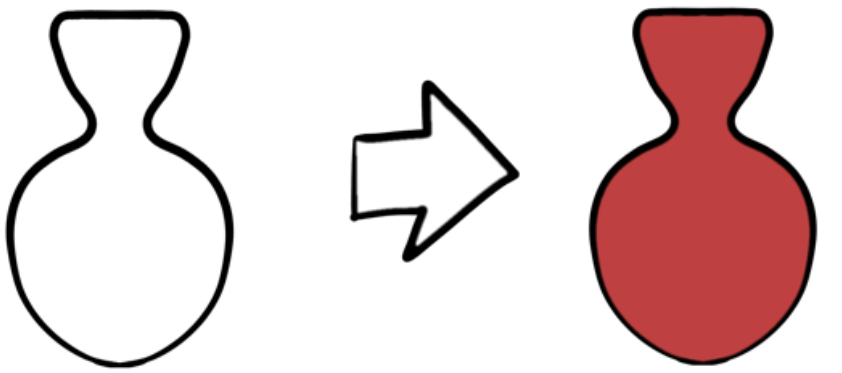
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.


My sharks!

Maybe we think impact is about change in particle positions,



Or maybe we think it's about change in object identities.



Vase

red vase

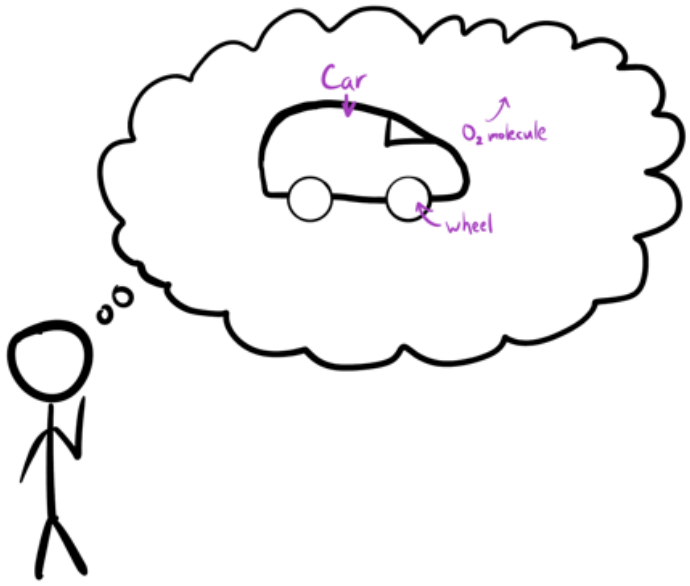
How can you not be tempted by that assumption –
things are changing in the world!
The assumption's right there.

It's so obvious.

And actually, it's totally wrong.

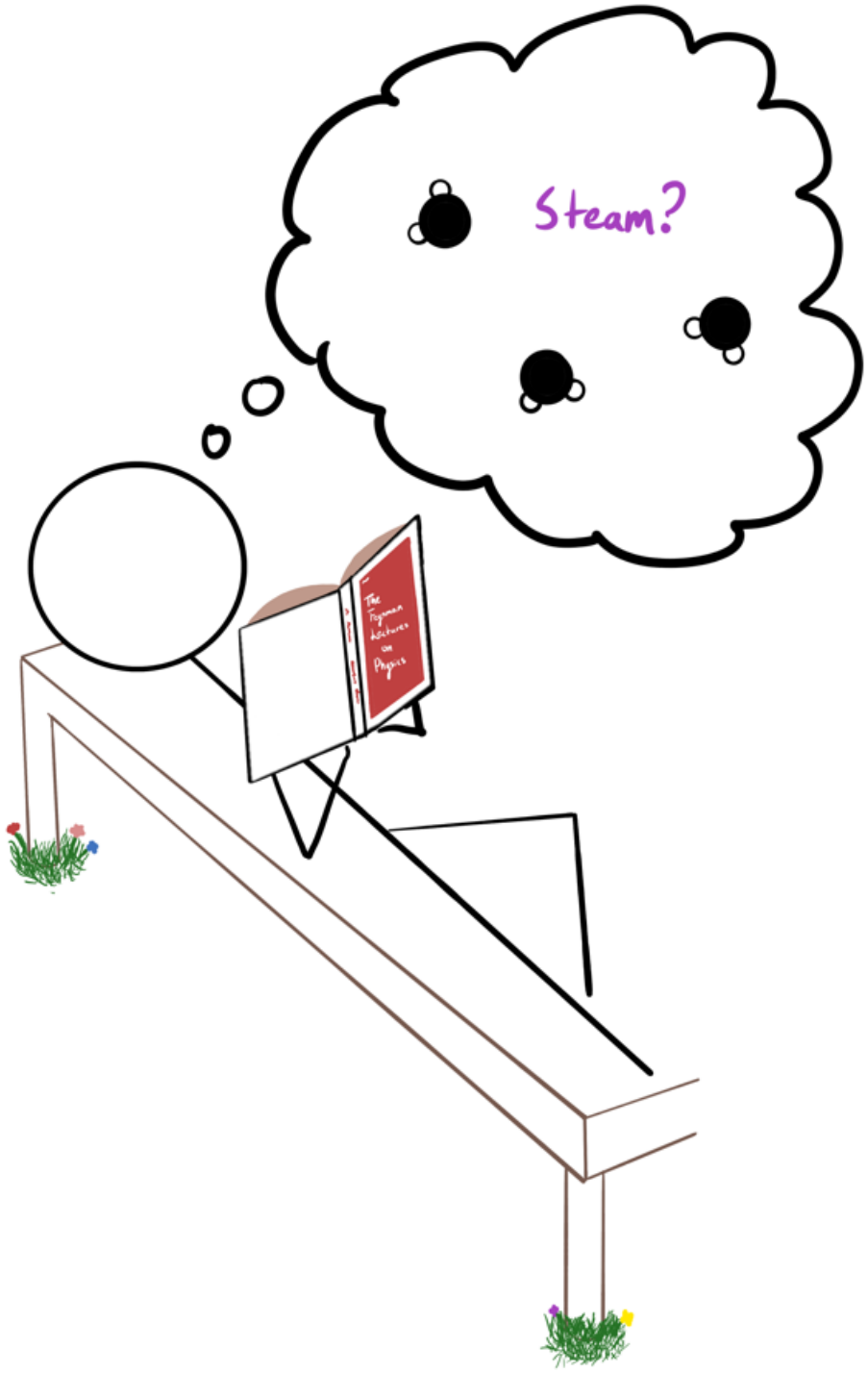
Impact is not Primarily about World State

Ontologies account for the things we think the world is made of.



Of course, we can think about cars and still know they're made of parts.

As we learn, we change our ontology.


Steam?
The Feynman Lectures on Physics

Your perceived AU is determined by the state of the world, but our ability to get what we want isn't usually affected by knowing about quarks. A calculator's output is determined by the state of the world, but the computation isn't about the state of the world.

So how do we know our intuitions are about AU, rather than a more direct function of the details of the world state? In the latter case, the function either does or doesn't change with our ontology.

If it does change with our ontology,
then ontological confusion would confuse our impact intuitions.

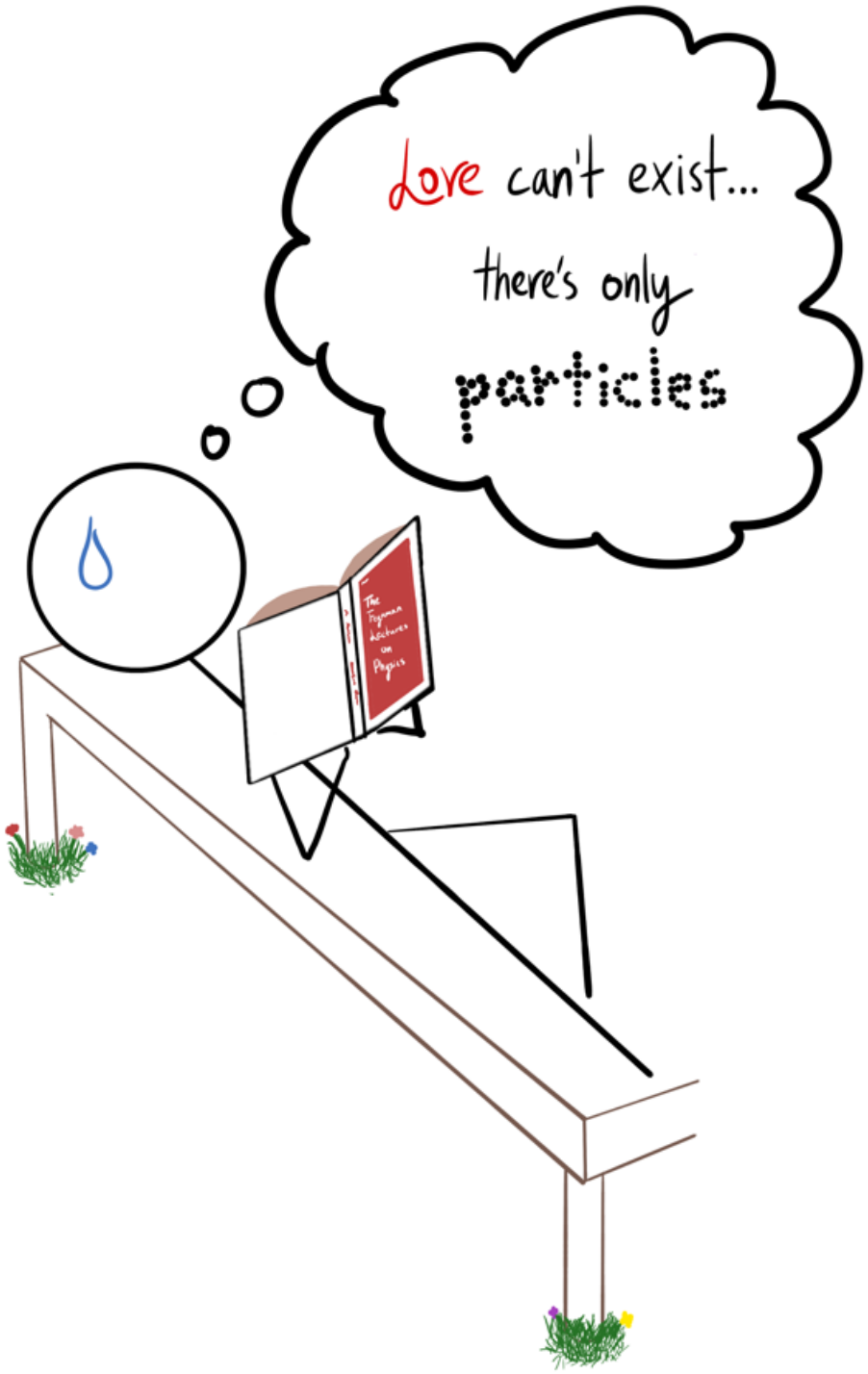
Pretend you have no idea what anything is made of, but you can still go about your day. You receive a $20,000 bonus.

How impactful does this feel?

Now pretend that you're back in your usual state of mind. Visualize receiving the bonus again; I predict it feels the same.

But perhaps the function doesn't change with the ontology.

When people think seriously about reductionism, they can have an ontological crisis.


Love can't exist...
there's only
particles
The Feynman Lectures on Physics

Usually, they realize valuable things can be made out of parts and get over it.

What happens to feelings of impact during an ontological crisis?

Fixed ontology theories: "nothing"

AU theory: "they become massively uncertain"

Two years ago, I had an ontological crisis. I vividly remember being unsure what was a big deal. When the crisis ended, impact went back to normal.

What's happening here is a failure to map our new representation of the world to things we find valuable.

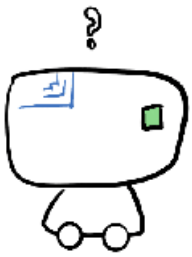
The impact uncertainty isn't because the ontology changes (we already ruled that out), but because we don't know whether there's any utility we want to attain!

Exercise: What happens to your intuitions about impact when you're unsure whether anything has value?

These existential crises also muddle our impact algorithm. This isn't what you'd see if impact were primarily about the world state.

Looking back, we see more evidence that impact doesn't hinge on an ontology.

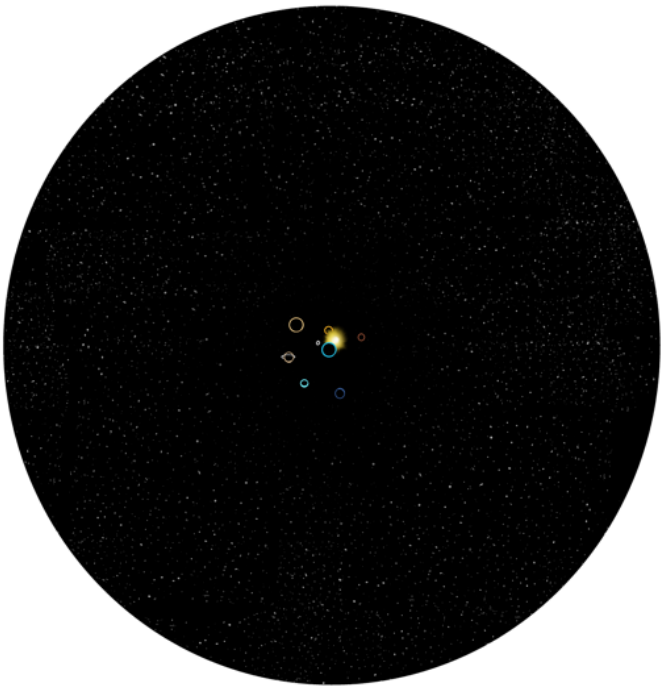
Did you stop to wonder how XYZ views the world?



Isn't it funny that the Pebblehoarders care so much about pebbles, while we care so much about suffering?

Remember the locality of objective impact?

Imagine I take a bunch of forever-inaccessible stars and jumble them up. This is a huge change in state, but it doesn't matter to us.



"How different is the world?"

is not the same mental question as

"How big of a deal is this?".

Impact is a thing that happens to agents.

Maybe you say,
"it's our ability to reach different world states weighted by how much we care"
– but that's just the AU theory with extra steps.

In fact, everything else seems to be extra steps.
Why is it important to preserve objects, or access to world states,
or anything else in the world?
Because of what they mean to us.

Conversely, why does attainable utility matter?
Does it reduce back to objects?

No. It matters because it matters;
the buck stops here, and seems to always stop here.

Imagine All theory came first and explained all these intuitions, and then someone suggests, "but what if we add in this stuff about the state?"

...

No thanks. After what we've seen, ontological theories shouldn't even be promoted to attention – that's privileging the hypothesis.

Let's consider what we now know:

- The locality of objective impact and the ontological confusion, ontological crisis, and existential crisis thought experiments are all evidence against ontological theories of impact.
- Every other consideration seems to just reduce to All.
- All theory is the simplest explanation.

I think that All theory isn't just an explanation, it's the explanation. Even if details are wrong, the correction won't produce a different kind of explanation.

Appendix: We Asked a Wrong Question

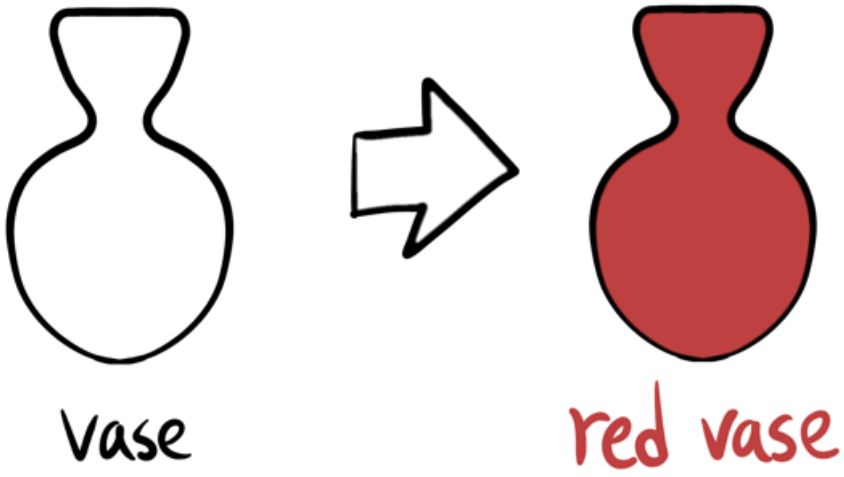
How did we go wrong?

> When you are faced with an unanswerable question—a question to which it seems impossible to even imagine an answer—there is a simple trick that can turn the question solvable.
>
> Asking “Why do I have free will?” or “Do I have free will?” sends you off thinking about tiny details of the laws of physics, so distant from the macroscopic level that you couldn’t begin to see them with the naked eye. And you’re asking “Why is X the case?” where X may not be coherent, let alone the case.
>
> “Why do I think I have free will?,” in contrast, is guaranteed answerable. You do, in fact, believe you have free will. This belief seems far more solid and graspable than the ephemerality of free will. And there is, in fact, some nice solid chain of cognitive cause and effect leading up to this belief.
>
> ~ Righting a Wrong Question

I think what gets you is asking the question "what things are impactful?" instead of "why do I think things are impactful?". Then, you substitute the easier-feeling question of "how different are these world states?". Your fate is sealed; you've anchored yourself on a Wrong Question.

At least, that's what I did.

Exercise: someone [me, early last year] *says that impact is closely related to change in object identities.*



Find at least two scenarios which score as low impact by this rule but as high impact by your intuition, or vice versa.

You have 3 minutes.

Gee, let's see... Losing your keys, the torture of humans on Iniron, being locked in a room, flunking a critical test in college, losing a significant portion of your episodic memory, ingesting a pill which makes you think murder is OK, changing your

discounting to be completely myopic, having your heart broken, getting really dizzy, losing your sight.

That's three minutes for me, at least (its length reflects how long I spent coming up with ways I had been wrong).

Appendix: Avoiding Side Effects

Some plans feel like they have unnecessary *side effects*:

Go to the store.

versus

Go to the store and run over a potted plant.

We talk about side effects when they affect our attainable utility (otherwise we don't notice), and they need both a goal ("side") and an ontology (discrete "effects").

Accounting for impact this way misses the point.

Yes, we can think about effects and facilitate academic communication more easily via this frame, but *we should be careful not to guide research from that frame*. This is why I avoided vase examples early on – their prevalence seems like a *symptom of an incorrect frame*.

(Of course, I certainly did my part to make them more prevalent, what with my first post about impact being called *Worrying about the Vase: Whitelisting*...)

Notes

- Your ontology can't be *ridiculous* ("everything is a single state"), but as long as it lets you represent what you care about, it's fine by AU theory.
- Read more about ontological crises at Rescuing the utility function.
- Obviously, something has to be physically different for events to feel impactful, but not all differences are impactful. Necessary, but not sufficient.
- AU theory avoids the mind projection fallacy; impact is subjectively objective *because* probability is subjectively objective.
- I'm not aware of others explicitly trying to deduce our native algorithm for impact. No one was claiming the ontological theories explain our intuitions, and they didn't have the same "is this a big deal?" question in mind. However, we need to actually understand the problem we're solving, and providing that understanding is one responsibility of an impact measure! Understanding our own intuitions is crucial not just for producing nice equations, but also for getting an intuition for what a "low-impact" Frank would do.

Misconceptions about continuous takeoff

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

There has been considerable debate over whether development in AI will experience a discontinuity, or whether it will follow a more continuous growth curve. Given the lack of consensus and the confusing, diverse terminology, it is natural to hypothesize that much of the debate is due to simple misunderstandings. Here, I seek to dissolve some misconceptions about the continuous perspective, based mostly on how I have seen people misinterpret it in my own experience.

First, we need to know what I even *mean* by continuous takeoff. When I say it, I mean a scenario where the development of competent, powerful AI follows a trajectory that is roughly in line with what we would have expected by extrapolating from past progress. That is, there is no point at which a single project lunges forward in development and creates an AI that is much more competent than any other project before it. This leads to the first clarification,

Continuous doesn't necessarily mean slow

The position I am calling "continuous" has been called a number of different names over the years. Many refer to it as "slow" or "soft." I think continuous is preferable to these terms because it focuses attention on the strategically relevant part of the question. It seems to matter less what the actual clock-time is from AGI to superintelligence, and instead matters more if there are will be single projects who break previous technological trends and gain capabilities that are highly unusual relative to the past.

Moreover, there are examples of rapid technological developments that I consider to be continuous. As an example, consider GANs. In 2014, GANs were used to generate low quality black-and-white photos of human faces. By late 2018, they were used to create nearly-photorealistic images of human faces.

Yet, at no point during this development did any project leap forward by a huge margin. Instead, every paper built upon the last one by making minor improvements and increasing the compute involved. Since these minor improvements nonetheless happened rapidly, the result is that the GANs followed a fast development relative to the lifetimes of humans.

Extrapolating from this progress, we can assume that GAN video generation will follow a similar trajectory, starting with simple low resolution clips, and gradually transitioning to the creation of HD videos. What would be unusual is if someone right now in late 2019 produces some HD videos using GANs.

Large power differentials can still happen in a continuous takeoff

Power differentials between nations, communities, and people are not unusual in the course of history. Therefore, the existence of a deep power differential caused by AI would not automatically imply that a discontinuity has occurred.

In a continuous takeoff, a single nation or corporation might still pull ahead in AI development by a big margin and use this to their strategic advantage. To see how, consider how technology in the industrial revolution was used by western European nations to conquer much of the world.

Nations rich enough to manufacture rifles maintained a large strategic advantage over those unable to. Despite this, the rifle did not experience any surprising developments which catapulted it to extreme usefulness, as far as I can tell. Instead, sharpshooting became gradually more accurate, with each decade producing slightly better rifles.

See also: Soft takeoff can still lead to decisive strategic advantage

Continuous takeoff doesn't require believing that ems will come first

This misconception seems to mostly be a historical remnant of the Hanson-Yudkowsky AI-Foom debate. In the old days, there weren't many people actively criticizing foom. So, if you disagreed with foom, it was probably because you were sympathetic to Hanson's views.

There are now many people who disagree with foom who don't take Hanson's side. Paul Christiano and AI Impacts appear somewhat at the forefront of this new view.

Recursive self-improvement is compatible with continuous takeoff

In my experience, recursive self improvement is one of the main reasons cited for why we should expect a discontinuity. The validity of this argument is far from simple, but needless to say: folks who subscribe to continuous takeoff aren't simply ignoring it.

Consider I.J. Good's initial elaboration of recursive self improvement,

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion’, and the intelligence of man would be left far behind.

The obvious interpretation from the continuous perspective is that by the time we have an ultraintelligent machine, we'll already have a not-quite-ultraintelligent machine. Therefore, the advantage that an ultraintelligent machine will have over the collective of humanity + machines will be modest.

It is sometimes argued that even if this advantage is modest, the growth curves will be exponential, and therefore a slight advantage right now will compound to become a large advantage over a long enough period of time. However, this argument by itself is not an argument against a continuous takeoff.

Exponential growth curves are common for macroeconomic growth, and therefore this argument should apply equally to any system which experiences a positive feedback loop. Furthermore, large strategic advantages do not automatically constitute a discontinuity since they can still happen even if no project surges forward suddenly.

Continuous takeoff is relevant to AI alignment

The misconception here is something along the lines of, "Well, we might not be able to agree about AI takeoff, but at least we can agree that AI safety is extremely valuable in either case." Unfortunately, the usefulness of many approaches to AI alignment appear to hinge quite a bit on continuous takeoff.

Consider the question of whether an AGI would defect during testing. The argument goes that an AI will have an instrumental reason to pretend to be aligned while weak, and then enter a treacherous turn when it is safe from modification. If this phenomenon ever occurs, there are two distinct approaches we can take to minimize potential harm.

First, we could apply extreme caution and try to ensure that no system will ever lie about its intentions. Second, we could more-or-less deal with systems which defect as they arise. For instance, during deployment we could notice that some systems are optimizing something different than what we intended during training, and therefore we shut them down.

The first approach is preferred if you think that there will be a rapid capability gain relative the rest of civilization. If we deploy an AI and it suddenly catapults to exceptional competence, then we don't really have a choice other than to get its values right the first time.

On the other hand, under a continuous takeoff, the second approach seems more promising. Each individual system won't by themselves carry more power than the sum of projects before it. Instead, AIs will only be slightly better than the ones that came before it, including any AIs we are using to monitor the newer ones. Therefore, to the extent that the second approach carries a risk, it will probably look less like a sudden world domination and will look more like a bad product rollout, in line with say, the release of Windows Vista.

Now, obviously there are important differences between current technological products and future AGIs. Still, the general strategy of "dealing with things as they come up" is much more viable under continuous takeoff. Therefore, if a continuous takeoff is more likely, we should focus our attention on questions which fundamentally can't be solved as they come up. This is a departure from the way that many have framed AI alignment in the past.

Introduction to Introduction to Category Theory

Category theory is so general in its application that it really feels like everyone, even non-mathematicians, ought to at least conceptually grok that it exists, like how everyone ought to understand the idea of the laws of physics even if they don't know what those laws *are*.

We expect educated people to know that the Earth is round and the Sun is big, even though those facts don't have any direct relevance to the lives of most people. I think people should know about Yoneda and adjunction in at least the same broad way people are aware of the existence and use of calculus.

But no one outside of mathematics and maybe programming/data science has heard of category theory, and I think a big part of that is because all of the examples in textbooks assume you already know what Sierpinski spaces and abelian groups are.

That is to say: all expositions of category theory assume you know math.

Which makes sense. Category theory is the mathematics *of math*. Trying to learn category theory without having most of an undergraduate education in math already under your belt is like trying to study Newton's laws without having ever seen an apple fall from a tree. You *can*...you're just going to have absolutely no intuition to rely on.

Category theory generalizes the things you do in the various fields of mathematics, just like how Newton's laws generalize the things you do when you toss a rock or push yourself off the ground. Except really, category theory generalizes what you do *when you generalize with Newton's laws*. Category theory *generalizes generalizing*.

Therefore, without knowing about any *specific generalizations*, like algebra or topology, it's hard to understand *general generalities*—which are categories.

As a result, there are no category theory texts (that I know of) that teach category theory to the educated and intelligent but mathematically ignorant person.

Which is a shame, because you totally can.

Sure, if you've never learned topology, plenty of standard examples will fly over your head. But every educated person has encountered the idea of generalization, and they've seen generalizations of generalizations. In fact, category theory is very intuitive, and I don't think it necessarily benefits from relating it all as quickly as possible to more familiar fields of mathematics. Instead, you should grasp the flow of category theory itself, as its own field.

So this is (tentatively, hopefully, unless I get busy, bored, or it just doesn't work out) a series on the basics of category theory *without assuming you know any math*. I'm thinking specifically of high school seniors.

There is no schedule for the posts. They'll just be up whenever I make them.

Why category theory? And why lesswrong?

Well, category theory is a super-general theory of everything. Rationality is also a super-general theory of everything. In fact, we'll see how category theory tells us a lot about what rationality really is, in a certain rigorous sense.

Basically...rationality comes from noticing certain general laws that seem to emerge every time you try to do something the "right" way. After a while, instead of focusing so much on the specifics, it starts being worth it to take a step back and study the general rules that seem to be emerging. And you start to notice that doing things the "right" way gets a lot easier when you *start* with the general rules and simply fill in the specifics, like how the quadratic formula makes quadratic equations a cinch to solve.

Category theory gives us *all* the general rules for doing things the "right" way.

(Don't actually hold me to demonstrating this claim.)

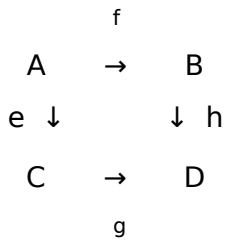
Why should *you* be interested in category theory?

One is because category theory is going to rise in importance in the future. It offers powerful new ways of doing math and science. So get started!

Two is that category theory makes it much easier to learn the rest of math. Well, *maybe*—this is an experiment, and a big motivation for doing this. How fast and well do people learn regular math if they can just say, "Oh, it's an adjunction" every time they learn a new concept?

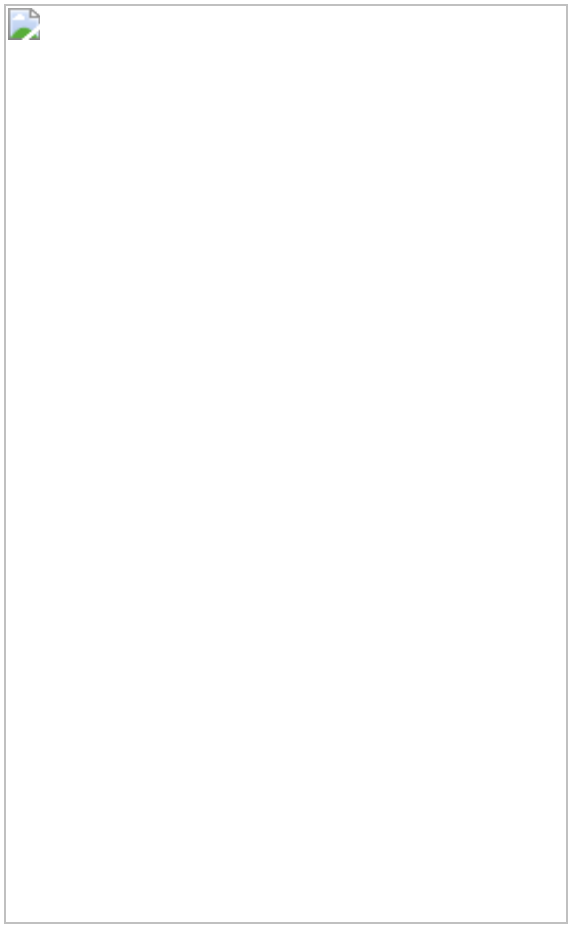
Three is that referencing homotopy type theory in conversation will make you sound cool and mysterious.

Please let me know if there's any interest in this.

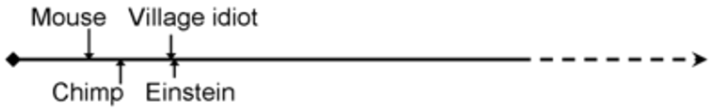


Two explanations for variation in human abilities

In My Childhood Role Model, Eliezer Yudkowsky argues that people often think about intelligence like this, with village idiot and chimps on the left, and Einstein on the right.



However, he says, this view is too narrow. All humans have nearly identical hardware. Therefore, the true range of variation looks something like this instead:



This alternative view has implications for an AI takeoff duration. If you imagine that AI will slowly crawl from village idiot to Einstein, then presumably we will have ample time to see powerful AI coming in advance. On the other hand, if the second view is correct, then the intelligence of computers is more likely to swoosh right past human level once it reaches the village idiot stage. Or as Nick Bostrom put it, "The train doesn't stop at Humanville Station."

Katja Grace disagrees, finding that there isn't much reason to believe in a small variation in human abilities. Her evidence comes from measuring human performance on various strategically relevant indicators: chess, go, checkers, physical manipulation, and Jeopardy.

In this post, I argue that the debate is partly based on a misunderstanding: in particular, a conflation of *learning ability* and *competence*. I further posit that when this distinction is unraveled, the remaining variation we observe isn't that surprising. Similar machines regularly demonstrate large variation in performance if some parts are broken, despite having nearly identical components.

These ideas are not original to me (see the Appendix). I have simply put two explanations together, which in my opinion, explain a large fraction of the observed variation.

Explanation 1: Distinguish learning from competence

Humans, despite our great differences in other regards, still mostly learn things the same way. We listen to lectures at roughly the same pace, we read at mostly the same speeds, and we process thoughts in similar sized chunks. To the extent that people *do* speed up lectures by 2x on Youtube, or speed read, they lose retention.

And putting aside tall tales of people learning quantum mechanics over a weekend, it turns out to be surprisingly difficult for humans to beat the strategy of long-term focused practice for becoming an expert at some task.

From this, we have an important insight: the range of learning abilities in humans is relatively small. There don't really seem to be humans who tower above us in terms of their ability to soak up new information and process it.

This prompts the following hypothetical objection,

> Are you sure? If I walk into a graduate-level mathematics course without taking the prerequisites seriously, then I'm going to be seriously behind the other students. While they learn the material, I will be struggling to understand even the most basic of concepts in the course.

Yes, but that's because you haven't taken the prerequisites. Doing well in a course is a product of learning ability * competence at connecting it to prior knowledge. If you

separate learning ability and competence, you will see that your learning ability is still quite similar to the other students.

Eric Drexler makes this distinction in Reframing Superintelligence,

> Since Good (1966), superhuman intelligence has been equated with superhuman intellectual competence, yet this definition misses what we mean by human intelligence. A child is considered intelligent because of learning capacity, not competence, while an expert is considered intelligent because of competence, not learning capacity. Learning capacity and competent performance are distinct characteristics in human beings, and are routinely separated in AI development. Distinguishing learning from competence is crucial to understanding both prospects for AI development and potential mechanisms for controlling superintelligent-level AI systems.

Let's consider this distinction in light of one of the cases that Katja pointed to above: chess. I find it easy to believe that Magnus Carlsen would beat me handily in chess, even while playing upside-down and blindfolded. But does this mean that Magnus Carlsen is vastly more intelligent than me?

Well, no, because Magnus Carlsen has spent several hours a day playing chess ever since he was 5, and I've spent roughly 0 hours a day. A fairer comparison is other human experts who have spent the same amount of practicing chess. And there, you might still see a lot of variation (see the next theory), but not nearly as much as between Magnus Carlson and me.

This theory can also be empirically tested. In my mind there are (at least) two formulations of the theory:

One version roughly states that, background knowledge and motivation levels being equal, humans will learn how to perform new tasks at roughly equal rates.

Another version of this theory roughly states that everything that top-humans can learn, most humans can too if they actually tried. That is, there *is* psychological unity of humankind in what we *can learn*, but not necessarily what we *have learned.* By contrast, a mouse really couldn't learn chess, even if they tried. And in turn, no human can learn to play 90-dimensional chess, unlike the hypothetical superintelligences that can.

You can also use this framework to cast the intelligence of machine learning systems in a new light. AlphaGo took a hundred million games to grow to the competence of Lee Sedol. But Lee Sedol has only played about 50,000 games in his life.

Point being, when we talk about AI getting more advanced, we're really mostly talking about what type of tasks computers can now learn, and how quickly. Hence the name *machine learning*...

Explanation 2: Similar architecture does not imply similar performance

Someone reading the above discussion might object, saying

> That seems right... but I still feel like there are some people who can't learn calculus, or learn to code no matter how hard they try. And I don't feel like this is just them not taking prerequisites seriously or lack of motivation. I feel like it's a fundamental limitation in their brain, and this produces a large amount of variation in learning ability.

I agree with this to an extent.[1] However, I think that there is a rather simple explanation for this phenomenon -- the same one in fact that Katja Grace pointed to in her article,

> Why should we not be surprised? [...] You can often make a machine worse by breaking a random piece, but this does not mean that the machine was easy to design or that you can make the machine better by adding a random piece. Similarly, levels of variation of cognitive performance in humans may tell us very little about the difficulty of making a human-level intelligence smarter.
>
> In the extreme case, we can observe that brain-dead humans often have very similar cognitive architectures. But this does not mean that it is easy to start from an AI at the level of a dead human and reach one at the level of a living human.

Imagine lining up 100 identical machines that produce widgets. At the beginning they all produce widgets at the same pace. But if I go to each machine and break a random part -- a different one for each machine -- some will stop working completely. Others will continue working, but will sometimes spontaneously fail. Others will slow down their performance, and produce fewer widgets. Others still will continue unimpeded since the broken part was unimportant. This is all despite the fact that the machines are nearly identical in design space!

If a human cannot learn calculus, even after trying their hardest, and putting in the hours, I would attribute this fact to a learning deficit. In other words, they have a 'broken part.' Learning deficits can be small things: I often procrastinate on Reddit rather than doing tasks that I should be doing. People's minds drift off when they're listening to lectures, finding the material to be boring.

The brightest people are the ones who can use the full capacity of their brains to learn the task. In other words, the baseline of human performance is set by people who have no 'broken parts.' Among those people, the people with no harmful mutations or cognitive deficits, I expect human learning ability to be extremely similar. Unfortunately there aren't really any humans who fit this description, and thus we see some variation, even for humans near the top.

Appendix

What about other theories?

There have been numerous proposed theories that I have seen. Here are a few posts:

Why so much variance in human intelligence? by Ben Pace. This question prompted the community to find reasons for the above phenomenon. None of the answers quite match my response, hence why I created this post. Still, I think a few of the replies were onto something and worth reading.

Where the Falling Einstein Meets the Rising Mouse, a post on SlateStarCodex. Scott Alexander introduces a few of his own theories, such as the idea that humans are simply lightyears above animals in abilities. While this theory seems plausible in some domains, I did not ultimately find it compelling. In light of the learning ability distinction, however, I *do* think we are lightyears above many animals in learning.

The range of human intelligence by Katja Grace. Her post prompted me to write this one. I agree that there is a large variation in human *abilities* but I felt that the lack of a coherent distinction between *learning* and *competence* ultimately made the argument rest on a confusion. As for whether this alters our perception of a more continuous takeoff, I am not so sure. I think the implications are non-obvious.

Where did the learning/competence distinction come from?

To the best of my knowledge, Eric Drexler kindled the idea. However, I do know that many people have made the distinction in the past; I just haven't really seen it applied to this debate in particular.

---

[1] However, I do think that most people are probably able to learn calculus, and how to code. I agree with Sal Khan who says,

> If we were to go 400 years into the past to Western Europe, which even then, was one of the more literate parts of the planet, you would see that about 15 percent of the population knew how to read. And I suspect that if you asked someone who did know how to read, say a member of the clergy, "What percentage of the population do you think is even capable of reading?" They might say, "Well, with a great education system, maybe 20 or 30 percent." But if you fast forward to today, we know that that prediction would have been wildly pessimistic, that pretty close to 100 percent of the population is capable of reading. But if I were to ask you a similar question: "What percentage of the population do you think is capable of truly mastering calculus, or understanding organic chemistry, or being able to contribute to cancer research?" A lot of you might say, "Well, with a great education system, maybe 20, 30 percent."

Edit: I want to point out that this optimism is *not* a crux for my main argument.

Effect heterogeneity and external validity in medicine

Our paper "Effect heterogeneity and variable selection for standardizing causal effects to a target population" has just been publised in the European Journal of Epidemiology at https://link.springer.com/article/10.1007/s10654-019-00571-w . While the journal's version of record is behind a paywall, a preprint is available on arXiv at https://arxiv.org/pdf/1610.00068.pdf.

This paper argues for my very deeply held belief that *we can make significant advances in quantitative reasoning for medical decision making by thinking more closely about effect heterogeneity and how this relates to the choice of effect scale.*

Over the course of the last 7 years, external validity and generalizability have become increasingly hot topics in statistical methodology and computer science. In particular, a lot of progress has been made by Judea Pearl and Elias Bareinboim, who introduced a framework based on causal diagrams that can be used to reason about how to take causal information from one setting (for example: a randomized trial) and apply it in a different setting (for example: a clinically relevant target population).

The key questions of interest are: How do we know whether such extrapolation is even possible? How do we determine what information we need from the study, and what information we need from the target population, in order to extrapolate the findings? How do we put this information together in order to obtain a valid prediction for what happens if the intervention is implemented in the target population?

Pearl and Bareinboim's framework for answering these questions is, of course, mathematically valid. However, in my opinion, their approach also throws the baby out with the bathwater. In particular, we argue that instead of attempting to extrapolate the magnitude of the effect (i.e. a measure of the "size" of the difference between what happens if the drug is taken, and what happens if the drug is not taken), they attempt to look at the people who were assigned to receive the intervention in the study, and extrapolate their distribution of outcomes to the target population *without any reference to how that distribution differs from what happened to the people who were randomized to the control condition*.

Theoretically, this approach will work if the extrapolation procedure can account for every cause of the outcome whose distribution differs between the people who were in the study, and the people you are trying to make predictions about. However, since the set of causes of the outcome is very large, it is very unlikely that it is possible to measure all of them. Moreover, it is very likely that we do not even know what the causes of the outcome are. Our inferences then become subject to potential uncertainty which arises from the auxiliary assumption that we know what to control for.

Consider a situation where scientists have conducted a randomized controlled trial in men, on the effects of homeopathy on heart disease. The scientists find that homeopathy has no effects in men, and wonder whether this finding can be extrapolated to women. If the scientists attempt to answer this question using Bareinboim and Pearl's framework, they will be forced to conclude that no

extrapolation can be made, unless they are willing to claim that they know all the causes of heart disease that differ between men and women, and have been able to measure every one of these causes in all the patients in the study.

In contrast, we suggest that scientists who want to extrapolate their findings and make predictions outside of the study should attempt to quantify the size of the effect - that is, by how much the outcomes in the people who were randomized to receive the intervention differ from the outcomes in the people who were randomized to the control condition. This effect size could then potentially be used as the basis for extrapolation. Such an approach would correspond much closer to how external validity and extrapolation has traditionally been understood in the medical literature.

In the real world of clinical medicine, doctors are usually given information about the effects of a drug on the risk ratio scale (the probability of the outcome if treated, divided by the probability of the outcome if untreated). With information on the risk ratio, a doctor may make a prediction for what will happen to the patient if treated, by multiplying the risk ratio and patient's risk if untreated (which is predicted informally based on observable markers for the patient's condition).

The problem with this approach is that there are multiple scales on which to quantify the magnitude of the effect. Other possible scales for measuring effects include:

- The odds ratio, which applies a $\frac{p}{1-p}$ transformation to the risk
- The survival ratio, which uses the probability of survival (1-p) instead of the probability of death (p)
- The risk difference (which uses an additive scale instead of a multiplicative one)

Unless the intervention has no effect, the empirical predictions will not be invariant to the choice of scale. This is, of course, a serious problem for principled clinical decision-making, but as we will show, it is not necessarily an impossible one.

Despite the scale dependence of the reasoning procedure, the risk ratio is in many cases the only summary of the effect size that is made available to clinicians, whether they get their information from journals, clinical guidelines or online resources for clinical information. Given that the reasoning procedure is not scale-invariant, the universal reliance on the risk ratio may plausibly lead to suboptimal medical decision making in a wide range of clinical scenarios. But, in contrast to the implications of the Bareinboim/Pearl framework, we argue that this does **not** necessarily mean that we should throw out reliance on parametric effect measures altogether.

Our suggestion for how to choose the scale has been discussed earlier on Less Wrong (see https://www.lesswrong.com/posts/K3d93AfFE5owfpkx4/counterfactual-outcome-state-transition-parameters). I am not going to repeat the argument in full here, but I will ask you to consider the following highly stylized thought experiment, which illustrates the underlying intuition:

Consider a randomized controlled trial where the intervention is that everyone is randomized to play Russian roulette once a year. This trial is conducted in Russia. It is found that among those who did not play Russian roulette, 1% of people died over the course of the year. Among the people who played Russian roulette, 18% of people died. We want to extrapolate these findings to Norway, where nobody ordinarily plays Russian roulette and it is known that 0.5% of people die during any year. Our goal is to

find out what happens in Norway if everyone took up playing Russian roulette once a year.

Bareinboim and Pearl would suggest taking the risk of death among those who played Russian roulette (18%), controlling for all causes of death that differ between Russia and Norway, and producing an estimate for what happens in Norway if everyone plays Russian roulette. However, due to considerable differences between Russia and Norway in terms of predictors of mortality, this is clearly not feasible in this situation.

If we instead attempt to quantify the effect size in Russia, this can be done on any of the previously discussed scales:

- The risk ratio is $\frac{0.18}{0.01} = 18$
- The risk difference is $0.18 - 0.01 = 0.17$
- The survival ratio is $\frac{1-0.18}{1-0.01} \approx \frac{5}{6}$
- The odds ratio is $\frac{\frac{0.18}{1-0.18}}{\frac{0.01}{1-0.01}} \approx 21.7$

Each of these scales will result in a different prediction for what will happen if people in Norway play Russian roulette:

- If we use the risk ratio, we will predict that $0.005 \times 18 = 9\%$ will die.
- If we use the risk difference, we will predict that $0.005 + 0.17 = 17.5\%$ will die.
- If we use the survival ratio, we will predict that $(1 - 0.005) \times \frac{5}{6} \approx 82.9\%$ will survive, meaning that 17.1% will die
- If we use the odds ratio, we will predict that $\frac{0.005 \times 21.7}{1 + 0.005 \times 21.7} \approx 9.8\%$ will die.

These predictions differ massively not only in their implications for decision-making but also in their plausibility: Given what we know about Russian roulette, we would expect to see results much closer to 17% than to 9%. So clearly, some of these scales are doing something "right" and other scales are doing something "wrong".

We argue that the key to understanding the implications of this scale-dependence is that only the survival ratio ($\frac{5}{6}$) has a structural meaning: it represents the proportion of empty chambers in the revolver, and therefore produces appropriate, valid predictions. In contrast, the risk ratio ($\frac{18}{1}$) has no possible structural meaning and therefore produces nonsense results.

Any attempt at extrapolation would, of course, have to account for all factors that determine the magnitude of the effect. For example, if Russians are more likely to be drunk when they play Russian roulette, they may be more likely to miss than Norwegians. This may lead to local deviations from effect sizes of $\frac{5}{6}$, which will have implications for extrapolation. But once you have controlled for all of the factors that determine the magnitude of the effect *on a scale that has structural meaning*, extrapolation may be valid.

Crucially, we argue that controlling for all determinants of effect size (alcohol? how many chambers are there in typical revolvers in each country?) is much more tractable than controlling for all causes of mortality differences between the countries.

The main idea behind my research agenda is to explore how far we can push this argument in more clinically relevant settings. Next, consider a doctor who is trying to determine the pros and cons of treating a patient with a new drug. Suppose a reliable study on the drug shows that among those who received a placebo, 1% got an allergic reaction over the following 12 months; whereas, among those who received the drug, 2% got an allergic reaction.

The scientists behind the study can either tell the doctor that the risk ratio is $\frac{0.02}{0.01} = 2$, or that the survival ratio is $\frac{0.98}{0.99} \approx 0.99$. Both statements are correct, but only the latter has a potential structural interpretation, since it plausibly corresponds to a state of nature where 99% of the population do not have the factors (genes?) that predispose a person to have an allergic reaction if exposed to the drug.

Now consider that this patient also has a severe peanut allergy (which is unrelated to the medical issues that the doctor is treating them for) and lives in an environment where everyone eats peanuts all the time. This patient, therefore, has a 10% baseline risk of getting an allergic reaction over the course of 12 months, even in the absence of treatment with the new drug.

It would be insanity for the doctor to expect that the risk ratio from the study generalizes, and that the patient will have a 20% risk of anaphylaxis if given the new drug. In contrast, it may be meaningful to predict that their risk under treatment is given by $1 - (1 - 0.10) \times \frac{0.98}{0.99} \approx 10.9\%$. This will correspond closely to what one might expect would happen if the patient belongs to a population that has the same distributions of factors that predispose to the specific drug-related allergic reaction, as the population that was studied in the trial.

For these reasons, I consider it crucial for medical scientists to become aware of the need to put significant effort into reasoning about whether an effect measure has plausible structural meaning in the context of their current research question, before deciding to use it as a summary of their findings which is suitable for use in clinical decision making.

If anyone can spot any flaws in our argument, such feedback would be invaluable information. I invoke Crocker's Rules for all responses to the paper and the post. I would very much appreciate it if this blog post and the paper could be forwarded to anyone who is in a position to evaluate its importance.

Finally, let me note that this paper is the first peer-reviewed academic publication to acknowledge support from the EA Hotel Blackpool in its funding section. The EA Hotel is a project worth supporting; see https://forum.effectivealtruism.org/posts/uyvc6p99vsWFMPZiz/ea-hotel-fundraiser-5-out-of-runway

How feasible is long-range forecasting?

This is a linkpost for https://www.openphilanthropy.org/blog/how-feasible-long-range-forecasting

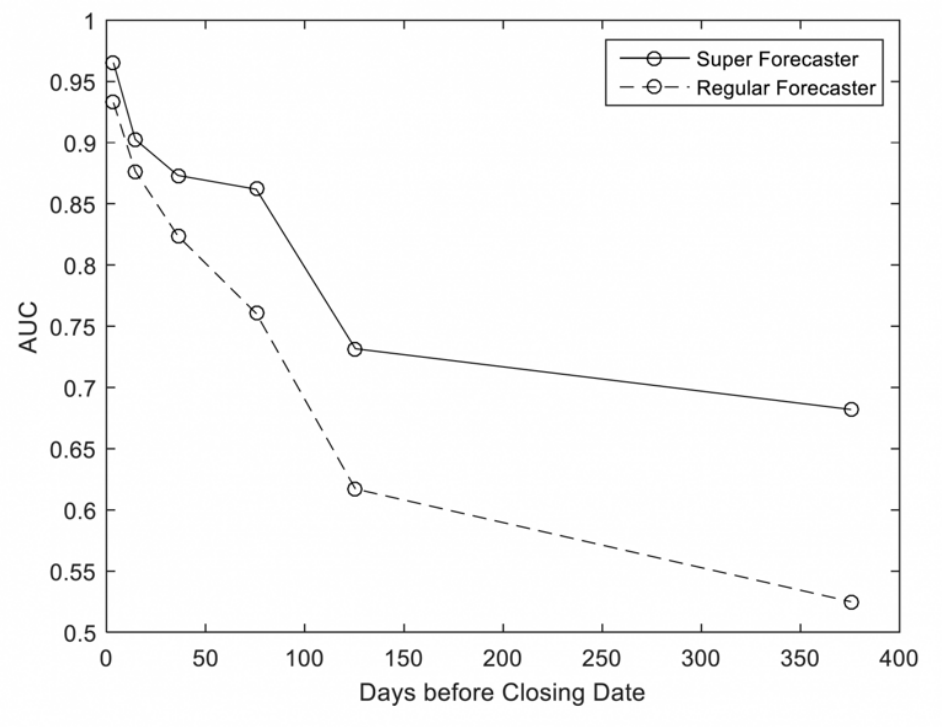
Lukeprog posted this today on the blog at OpenPhil. I've quoted the opening section, and footnote 17 which has an interesting graph I haven't seen before.

How accurate do long-range (≥10yr) forecasts tend to be, and how much should we rely on them?

As an initial exploration of this question, I sought to study the track record of long-range forecasting exercises from the past. Unfortunately, my key finding so far is that it is difficult to learn much of value from those exercises, for the following reasons:

1. Long-range forecasts are often stated too imprecisely to be judged for accuracy.
2. Even if a forecast is stated precisely, it might be difficult to find the information needed to check the forecast for accuracy.
3. Degrees of confidence for long-range forecasts are rarely quantified.
4. In most cases, no comparison to a "baseline method" or "null model" is possible, which makes it difficult to assess how easy or difficult the original forecasts were.
5. Incentives for forecaster accuracy are usually unclear or weak.
6. Very few studies have been designed so as to allow confident inference about which factors contributed to forecasting accuracy.
7. It's difficult to know how comparable past forecasting exercises are to the forecasting we do for grantmaking purposes, e.g. because the forecasts we make are of a different type, and because the forecasting training and methods we use are different.

Despite this, I think we can learn *a little* from GJP about the feasibility of long-range forecasting. Good Judgment Project's Year 4 annual report to IARPA (unpublished), titled "Exploring the Optimal Forecasting Frontier," examines forecasting accuracy as a function of forecasting horizon in this figure (reproduced with permission):



This chart uses an accuracy statistic known as AUC/ROC (see Steyvers et al. 2014) to represent the accuracy of binary, non-conditional forecasts, at different time horizons, throughout years 2-4 of GJP. Roughly speaking, this chart addresses the question: “At different forecasting horizons, how often (on average) were forecasters on ‘the right side of maybe’ (i.e. above 50% confidence in the binary option that turned out to be correct), where 0.5 represents ‘no better than chance’ and 1 represents ‘always on the right side of maybe’?”

For our purposes here, the key results shown above are, roughly speaking, that (1) regular forecasters did approximately no better than chance on this metric at ~375 days before each question closed, (2) superforecasters did substantially better than chance on this metric at ~375 days before each question closed, (3) both regular forecasters and superforecasters were almost always “on the right side of maybe” immediately before each question closed, and (4) superforecasters were roughly as accurate on this metric at ~125 days before each question closed as they were at ~375 days before each question closed.

If GJP had involved questions with substantially longer time horizons, how quickly would superforecaster accuracy declined with longer time horizons? We can’t know, but an extrapolation of the results above is at least compatible with an answer of “fairly slowly.”

I'd be interested to hear others' thoughts on the general question, and any opinions on the linked piece.

Implementing an Idea-Management System

This post is for you if:

1. Projects that excite you are growing to be a burden on your to-do list
2. You have a nagging sense that you're not making the most of the ideas you have every day
3. Your note- and idea-system has grown to be an unwieldy beast

Years ago, I ready David Allen's "Getting Things Done". One of the core ideas is to write down everything, collect it in an inbox and sort it once a day.

This lead to me writing down tons of small tasks. I used Todoist to construct a system that worked for me — and rarely missed tasks.

It also lead to me getting a lot of ideas, that could sprout a bunch of tasks. But ideas are much different from tasks to me. They're something that I might, but probably won't complete. Something that may come in useful in the future. And plain fun to come up with.

I stored them in Todoist as well, but recently I've started considering whether that's wise. They started weighing on me. It became a growing list of possibilities, many of which I'd never finish. The task at the top of my list became the top of my priorities, simply because of its location.

There must be a better way.

But what might that look like?

The ideal idea-management system

1. Separates ideas from commitments

I want a system that separates obligations from ideas. Form follows function, so I'd prefer something that doesn't structure ideas in lists. This rules out Todoist completely.

2. Shows you the right ideas at the right time

Even with complete foreknowledge, finding the perfect schedule might be practically impossible. In contrast, thinking on your feet and reacting as jobs come in won't give you as perfect a schedule as if you'd seen into the future-but the best you can do is much easier to compute. — Algorithms to Live By

Most of us live dynamic lives where priorities change often. Your children start a new hobby, you're handed a task at work, or you can finally work on your passion-project. Your idea-management system should reflect this. It shouldn't just show you your

most recent idea, it should make it easy to find ideas associated with whatever you find most important right now.

Avoiding lists makes it more likely that you take action on ideas that matter to you right now. You don’t skim from the top, you go for the area that matters and find ideas related to it.

If you sort ideas around a central node, you can pin-point synergies and conflicts. If you’ve taken notes on 4 different project-management systems, you want to see them all when you need them.

3. Doesn’t distract you with ideas that you can’t execute

You don’t want to waste time considering ideas that aren’t important right now. Sometimes you’re missing resources, or you’re waiting for some dependency.

Project/idea-lists are terrible at this. As you skim through them, a plethora of memories activate, most of which are irrelevant to what you end up doing.

4. Allows you to break down ideas into smaller parts, and re-combine them as needed

Tiago Forte’s Intermediate Packets inspired this. Many of our ideas can work in exactly the same way.

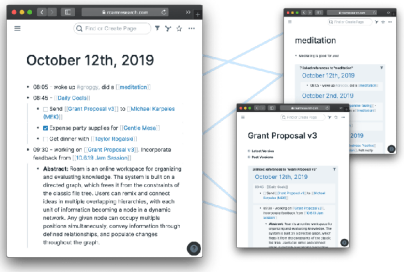
5. Has low overhead

You want to spend as little time as possible sorting and searching through your ideas. This should be a no-brainer. You want it to be easy to inter-link ideas and to add reference material. And when you get an idea on the train, you want to offload it without wasting time.

Time for action

Okay, Martin, I’m sold. But folders, task-managers, outliners like Workflowy and Dynalist — they’re all hierarchical!

You’re right, and until last week, I didn’t know what other solution there could be. But now there is. Roam.



Roam is different. It makes it trivial to link- and back-link pages. It creates a new page just by linking to it. And it back-links as well! When you link [[Self-determination theory]] to [[Motivation]], the motivation page will show a link to Self-determination theory in its footnotes.

This is tremendous. It creates a clear divide between commitments and ideas. Commitments belong on lists, ideas in dynamic networks.

When you get a new idea on a motivation tweak, you add it and link it to [[Motivation]].

Most of the time, everything is going well, so you don't need it right now. But 6 months later, you're assigned a grind of a task. You decide to read up on on Motivation, and voilá, in the footnotes is a link to that idea you had that might help you now.

Not only that, all your other ideas on motivation are there, for you to synergise or compare/contrast.

And you're less distracted. You don't have to take action on an idea in fear of forgetting it. Nor are you presented with ideas only because they're recent. You have a need, [[Motivation]], and you're presented with ideas on that topic alone. No distraction.

Roam also allows you to link to any bullet-point in any other note. Say you want your collaborators to identify with the core values of your projects. Why not embed that idea you encountered 3 months ago from Organismic Integration Theory? In this way, you can re-use sub-ideas from other major themes in any of your other projects. And if you find out that idea didn't work? You add a note, and that note propagates to any other places you've referenced the idea.

Roam becomes your second brain. You draw associations, and Roam remembers. You want to look something up, and Roam shows you what you've considered relevant in the past.

How do you avoid losing track of important projects?

I advocate for using Roam as an idea-management system, not a project-management system. If you have an obligation, by all means track it in a list-style way that you review.

But if it’s an idea, you don’t want to spend time thinking about it when you’re executing something else. You want focus and produce, and to save the idea for when there’s time and a need. Don’t just be efficient, be effective.

There is nothing so useless as doing efficiently what should not be done at all — Peter Drucker

The Nitty Gritty, A Recipe for Implementation

For each idea that may turn into a project, I create a new page in this format:

“PI: Description”, eg. “PI: Research How to Effectively Integrate Motivations”

I have 3 prefixes:

1. PI for “Project Idea”
2. WO for “Working On”
3. AP for “Archived Project”

In each of these pages, I spend ~5 seconds referencing concepts where I may want to encounter the project. For this one, [[Motivation]], [[Self-Determination Theory]] and [[Organismic Integration Theory (OIT)]].

I also add any references I may want to read, and whichever ideas I’ve already had about the project.

This makes it trivial to pick up the idea when I have the time and need, and to execute it efficiently.

~

These posts are about getting ideas into the wild, having other people criticise them, making them better and connecting with like-minded people. So feel free to let me know what you think 🙂

I appreciate your time.

Originally published at http://martinbern.org on October 18, 2019.

We tend to forget complicated things

One consistent pattern I've noticed in studying math is that, if some material feels very difficult, then I might remember it in an upcoming exam, but I will almost certainly have forgotten most of it one year later. The success story behind permanent knowledge gain is almost always "this was hard once but now it's easy, so obviously I didn't forget it" and almost never "I successfully memorized a lot of complicated-feeling things."

I think this also applies outside of mathematics. If it's roughly correct, then the most obvious consequence is to adapt your behavior when you're learning something. Provided that your goal is to improve your understanding permanently by understanding the material conceptually (which, of course, may not be the case), either study until it gets easy, or decide it's not worth your time at all, but don't stop when you've just barely understood it.

I've violated this rule many times, and I think it has resulted in some pretty inefficient use of time.

How do you assess the quality / reliability of a scientific study?

When you look at a paper, what signs cause you to take it seriously? What signs cause you to discard the study as too poorly designed to be much evidence one way or the other?

I'm hoping to compile a repository of heuristics on study evaluation, and would love to hear people's tips and tricks, or their full evaluation-process.

I'm looking for things like...

- "If the n (sample size) is below [some threshold value], I usually don't pay much attention."
- "I'm mostly on the lookout for big effect sizes."
- "I read the abstract, then I spend a few minutes thinking about how I would design the experiment, including which confounds I would have to control for, and how I could do that. Then I read the methods section, and see how their study design compares to my 1-3 minute sketch. Does their design seem sensible? Are they accounting for the first-order-obvious confounds?"
- etc.

Gradient hacking

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

"Gradient hacking" is a term I've been using recently to describe the phenomenon wherein a deceptively aligned mesa-optimizer might be able to purposefully act in ways which cause gradient descent to update it in a particular way. In Risks from Learned Optimization, we included the following footnote to reflect this possibility:

> Furthermore, a deceptively aligned mesa-optimizer would be incentivized to cause there to be a systematic bias in the direction of preventing the base optimizer from modifying its mesa-objective. Thus, in the context of a local optimization process, a deceptive mesa-optimizer might try to “hack” its own gradient (by, for example, making itself more brittle in the case where its objective gets changed) to ensure that the base optimizer adjusts it in such a way that leaves its mesa-objective untouched.

Outside of that footnote, however, there hasn't been much other discussion of this phenomenon, and I think that's a shame, as I think it's definitely something that's worth looking into.

For example, here's a gradient hacking story that I'm particularly concerned about. Suppose you're training a model and it becomes deceptive. You're not worried about that, though, because you have sufficient transparency tools that you're confident you can detect any deceptive cognition it might be doing. If it's thinking about how to deceive you, your transparency tools will pick that up, and you'll know to throw out that model and try again. However, your transparency tools are quite expensive, so you can only afford to run them once, which you're planning on doing after training before you deploy your model. If your model is deceptive, though, then it might know all of that—in particular, it might know that if it stays deceptive, it'll get caught.

However, there is still something it can do in that situation to be able to defect against you and cause whatever it wants to happen during deployment. That is, it could hack its own gradient to produce a new model which will still exhibit the same deceptive behavior, but without doing any of the detectable deceptive cognition. Perhaps the new model exhibits the deceptive behavior not because it's trying to act deceptive, but just because it's a hard-coded part of its algorithm, or because it has some bad proxy, or any other defect that might be more difficult to detect than outright deception.

How might such a deceptive model accomplish such a feat of gradient hacking? Well, I don't think it's actually that difficult once your deceptive model comes up with it as a strategy. First, note that since gradient descent is a local optimization process, it doesn't have to somehow change what the global minima might be, or even what the local minima are—it just has to create a ridge in the loss landscape that will push the gradient in the direction it wants.

The original footnote provides one example of this, which is for the model to check if its objective satisfies some criterion, and fail hard if it doesn't. Now, if the model gets to the point where it's actually just failing because of this, then gradient descent will probably just remove that check—but the trick is never to actually get there. By having such a check in the first place, the model makes it so that gradient descent

won't actually change its objective, since any change to its objective (keeping all the other parameters fixed, which is what gradient descent does since it computes partial derivatives) would lead to such a failure.

The case of a deceptive model trying to erase evidence of its deception is not quite as straightforward, though I think it would still be possible to do something quite similar in that situation as well. This is especially true if the model has the ability to read and write to some sort of state (e.g. an RNN or even a transformer which can attend to its previous output). For example, it could write out a pre-commitment to follow the hard-coded policy and then just have a policy of always following any pre-commitments it finds such that it never thinks about the deception that led it to that policy in the first place.

So, that's the problem—how do we solve it? Well, there's a simple solution in the case above, which is just to run your transparency tools throughout training—not just at the end—so that you can catch the deception as soon as it happens. I think this is the right solution,[1] but it's worth noting that this adds an additional desideratum to your transparency tools, which is that they have to be cheap enough to run during the entire training process. That might not be that bad, however, if you don't do very many gradient descent update steps, and instead just use massive models and huge batch sizes—and there is some evidence that such a procedure is actually the most compute-efficient way to train neural networks. That being said, I think there are some transparency techniques which lend themselves better to being applied across training than others. Thus, the possibility of gradient hacking suggests that we should focus more on those techniques which allow us to not just understand a single model, but also the entire history of how that model was produced.[2]

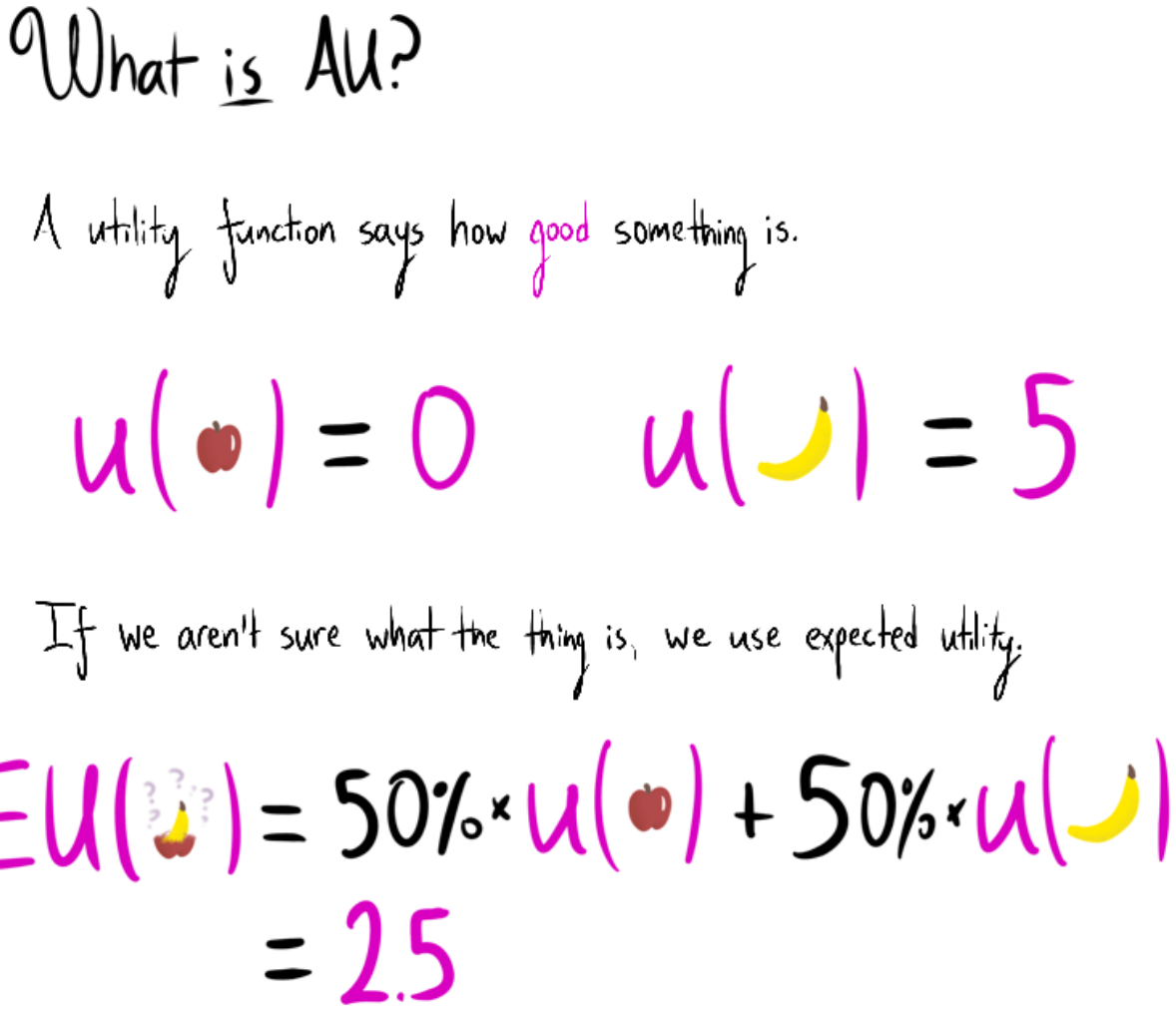
---

1. This is one of the reasons I'm interested in optimization provenance, for example. ↩

2. As an example of a technique which I think translates well to this domain, one of the OpenAI clarity team's current major research thrusts is "model diffing," which is about trying to understand exactly how two similar models differ from each other. Compared to other approaches that are just focused on understanding a single model, model diffing could be quite useful in allowing one to understand the entirety of the training process, since you could diff the model at each gradient descent update step. ↩

The Gears of Impact

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



What about attainable utility?

People have a natural sense of what they "could" do. If you're sad, it still feels like you "could" do a ton of work anyways. It doesn't feel physically impossible.

However, the part of you that predicts stuff doesn't buy that's gonna happen.

Beliefs about Future Actions

Imagine suddenly becoming not-sad.

Now, you "could" work when you're sad, and you "could" work when you're not-sad, so if AU just compared the things you "could" do, you wouldn't feel impact here.

But you did feel impact, didn't you?

So not only is AU not using the "could" algorithm, but it uses accurate predictions about how states of mind affect what you'll do later, and whether those actions will get you what you want.

AU ≈ "Embedded Agentic" EU

It seems like we have beliefs about our future actions.

Imagine having the following beliefs,

Where thinking more won't change the numbers:

Moping and watching Netflix

90% chance

Texting an ex you shouldn't

6% chance

Actually working

4% chance

What does your AU feel like?

Imagine you learned that this is the "real" distribution: you are, in fact, 90% likely to mope. Does learning this feel impactful?

For me, it doesn't feel like the AU has actually changed.

If AU relied even a bit on aspirational "could" actions, then learning they wouldn't come about would feel impactful.

AU seems to simply use our expectations.

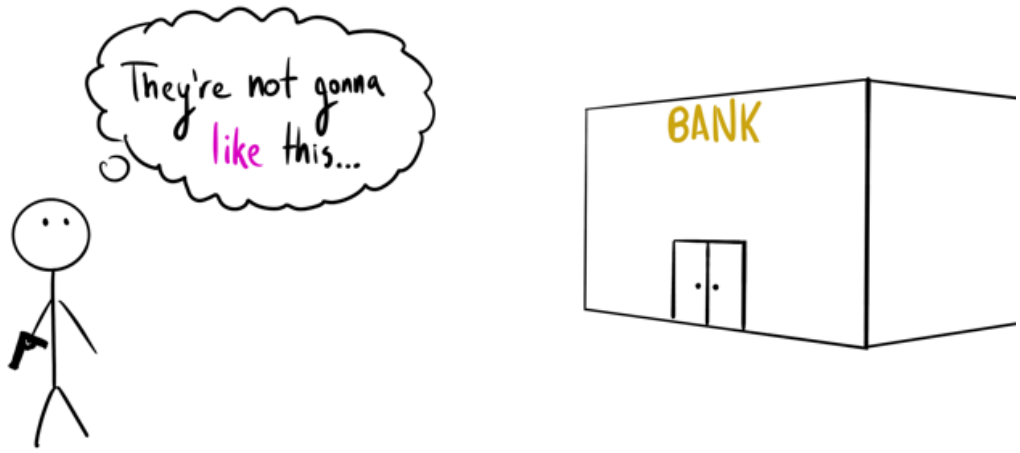
Conservation of Expected AU

Your beliefs should account for what you already know.

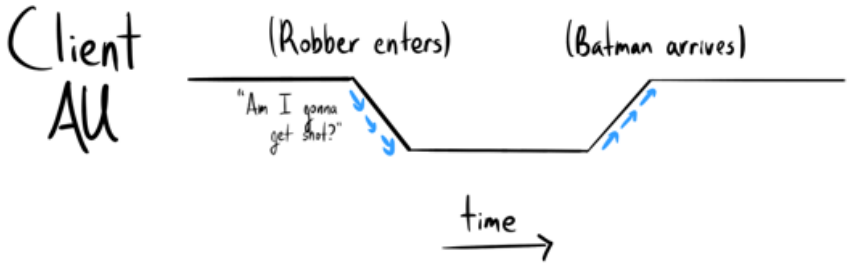
If you believe that tomorrow you'll have good reason to expect to get an A on your final exam, then that should change your mind today.

In the exact same way, your AU should account for what you already know - you don't expect to believe you can get a promotion without already believing it.

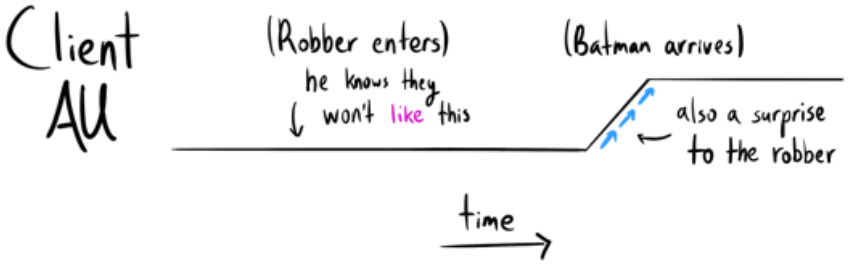
However, sometimes we can predict impact to others.



From the perspective of the bank's clients, it feels like



When the robber considers the clients' AU, it feels like



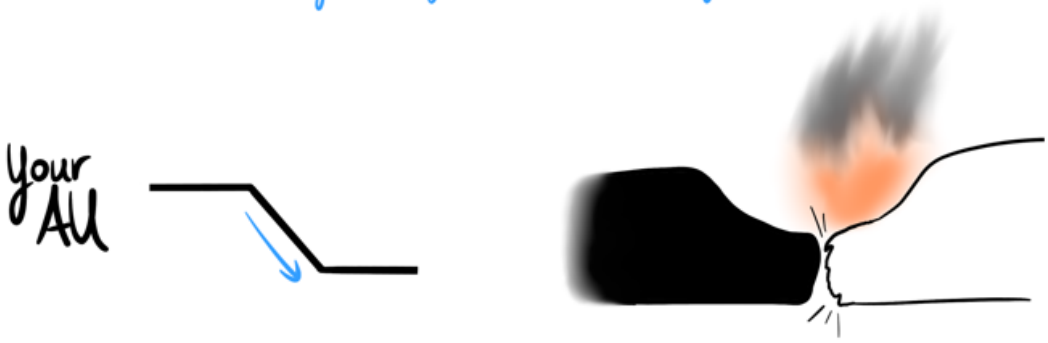
Let's switch gears a bit; we're now ready to understand

The Gears of Impact

How does a car crash affect you?

Which consequences are important to you, exactly?

Changes in your attainable utility, of course.

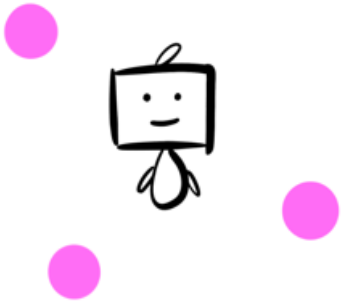


An event impacts us exactly and only when it changes what we think can happen for us. Probability mass has to shift.

		Before		After
Best	Go on a date	80%		1%
	Stuck in traffic	20%	→	0%
Worst	Paralyzed in hospital	0%		99%

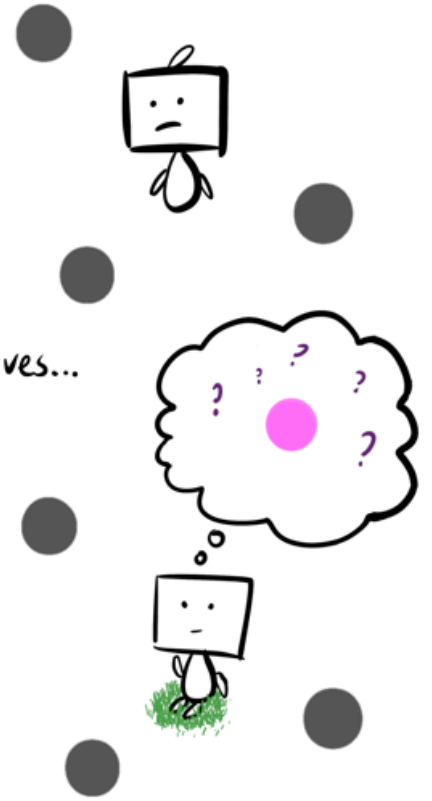
The degree to which important outcomes are affected by the event is the degree of the impact.

At this point, this might feel obvious, but don't forget how far we've come! We can think about this process with the Frank analogy.



Frank knows of pink objects, but doesn't think he can get to them.

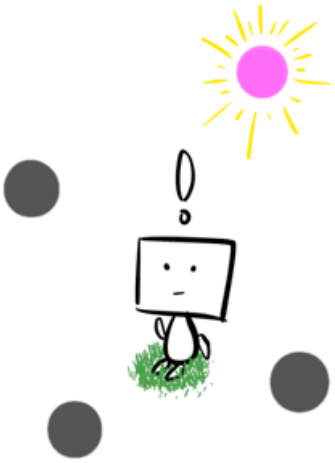
You've got plans tonight, but your friend flakes.



Frank considers alternatives...

You think about what you could do instead...

and finds one!



and remember another friend is available!

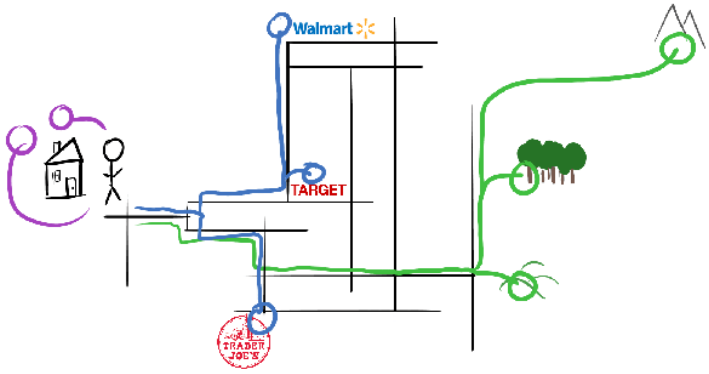
Once promoted to your attention, you see that the new plan isn't so much worse after all. The impact vanishes.

However, if you don't see a better alternative for some time, this becomes the new normal. If you then find a better alternative, this feels like a positive impact.

So, negative impact knocks out all of the best possibilities.

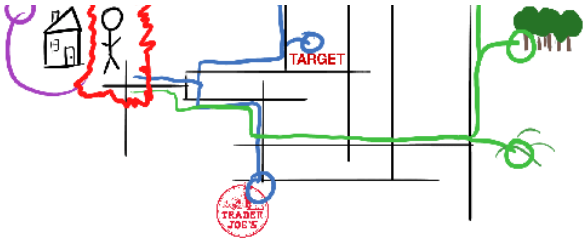
Consider the possibilities for different goals:

Relax in backyard Buy groceries Hike



Negative objective impact decreases your ability to achieve most goals.

But what could possibly eliminate or degrade all of these possibilities?

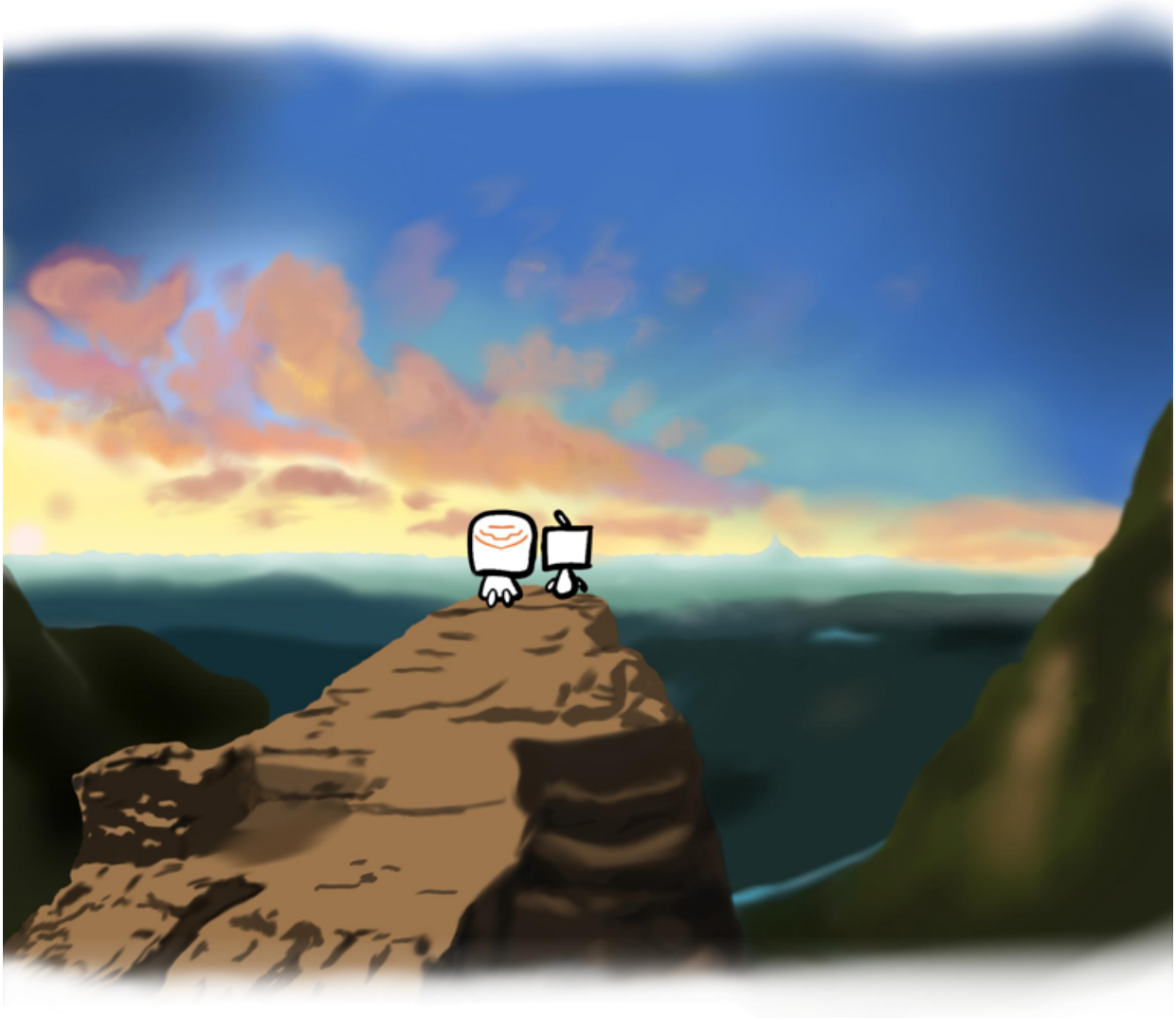


You are the common denominator.

Objective impact involves harm to you or your resources,
and this is why.

The first third of the sequence meets its close.
We understood why some things seem like big deals,
righted a wrong question, and just now
skirted the fascinating deeper nature of objective impact.

Objective impact, instrumental convergence, opportunity cost,
the colloquial meaning of "power" — these all prove to be
facets of one phenomenon, one structure.



Scheduling: The remainder of the sequence will be released after some delay.

Exercise: Why does instrumental convergence happen? Would it be coherent to imagine a reality without it?

Notes

- Here, our descriptive theory relies on our ability to have reasonable beliefs about what we'll do, and how things in the world will affect our later decision-making process. No one knows how to formalize that kind of reasoning, so I'm leaving it a black box: we *somehow* have these reasonable beliefs which are *apparently* used to calculate AU.
- In technical terms, AU calculated with the "could" criterion would be closer to an optimal value function, while actual AU seems to be an on-policy prediction, *whatever that means* in the embedded context. Felt impact corresponds to TD error.
    - This is one major reason I'm disambiguating between AU and EU; in the non-embedded context. In reinforcement learning, AU is a very particular

kind of EU: $V^*(s)$, the expected return under the optimal policy.

- Framed as a kind of EU, we plausibly use AU to make decisions.
- I'm not claiming normatively that "embedded agentic" EU *should* be AU; I'm simply using "embedded agentic" as an adjective.

Why are people so bad at dating?

I'm confused why people are so bad at dating. It seems to me like there are tons of $20 bills lying on the ground which no one picks up.

For example, we know that people systematically choose unattractive images for their dating profiles. Sites like PhotoFeeler cheaply (in some cases, freely) resolve this problem. Since photo quality is one of the strongest predictors of number of matches, you would think people would be clamoring to use these sites. And yet, not many people use them.

In the off-line dating world, it surprises me how few self-help books are about dating. Right now, zero of the top 10 Amazon best-selling self-help books are about dating. I see only two dating books in the top 50: The 5 Love Languages and Super Attractor. To the extent these books exist, they often have little to no empirical support; my guess is that horoscopes are the most frequently read source of dating advice. Evidence-based books like Mate are less widely read.

Possible Solution #1: Inadequate Equilibria

It might be that we are in an Inadequate Equilibrium. Eliezer proposes three general ways in which seeming inefficiencies can exist:

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

This doesn't seem very compelling in the case of online dating. Anyone could choose to use PhotoFeeler for themselves, for example.

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

Again, this isn't compelling. PhotoFeeler clearly lets you know what other people think of your photos.

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

Regressions done by Hitsch et al., as well as common sense, indicate that improving your own photos, even if you do nothing else or nothing else changes about the world, does make a significant impact in your likelihood of finding a good partner. So again, this seems uncompelling.

Possible Solution #2: Free Energy

> I've seen a number of novice rationalists committing what I shall term the Free Energy Fallacy, which is something along the lines of, "This system's purpose is

> supposed to be to cook omelettes, and yet it produces terrible omelettes. So why don't I use my amazing skills to cook some better omelettes and take over?"
>
> And generally the answer is that maybe the system from your perspective is broken, but everyone within the system is intensely competing along other dimensions and you can't keep up with that competition. They're all chasing whatever things people in that system actually pursue—instead of the lost purposes they wistfully remember, but don't have a chance to pursue because it would be career suicide. You won't become competitive along those dimensions just by cooking better omelettes. – An Equilibrium of No Free Energy

It's possible that people don't actually want to find good mates. Maybe they just want to *seem as though* they are trying to find good mates, or something. This would be consistent with dating advice being so evidence-free: people really want to signal that they care about finding good mate (which they can do by leaving a copy of Cosmo conspicuously out on their coffee table), but don't actually care about finding a good mate (so they don't care if Cosmo actually has good advice).

I'm pretty skeptical of this. If I was forced to guess only one thing that humans actually, really, really really really, valued as a terminal goal, "find a good mate" would be pretty high on my list of guesses. It's the thing we have millions of years of evolutionary pressure towards prioritizing. I might even go so far as to suggest that all the other markets which are efficient are efficient largely because of people's desire for romantic success: quants find arbitrage opportunities in the stock market because they hope that this financial success will translate into romantic success, etc.

So why is it that people – including people who devote their lives to finding arbitrage opportunities – leave so many metaphorical $20 bills on the ground when they start dating?

I remain confused.

When is pair-programming superior to regular programming?

When doing software development there are many choices that can affect your productivity. There are choices about how a company works. One of those is doing pair-programming over doing solo-programming. We recently had a consultant at our company who claimed that pair-programming is proven to be better but who couldn't point to evidence.

How does the case for pair-programming look like? What do we know? How can we rationally think about the question?

All I know is Goodhart

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

I've done some work on Goodhart's law, and I've argued that we can make use of all our known uncertainties in order to reduce or remove this effect.

Here I'll look at a very simple case: where we know only one thing, which is that Goodhart's law exists.

Knowing about Goodhart's law

Proxies exist

There are two versions of Goodhart's law, as we commonly use the term. The simplest is that there is a difference between maximising for a proxy -- V -- rather than for the real objective -- U.

Let $W = U - V$ be the difference between the true objective and the proxy. Note that in this post, we're seeing U, V, and W as actual maps from world histories to $\mathbb{R}$. The equivalence classes of them under positive affine transformations are denoted by $[U]$, $[V]$, $[W]$ and so on.

Note that $U = V + W$ makes sense, as does $[U] = [V + W]$, but $[U] = [V] + [W]$ does not: $[V] + [W]$ defines a family of functions with three degrees of freedom (two scalings and one addition), not the usual two for $[U]$ (one scaling and one addition).

So, the simplest version of Goodhart's law is thus that there is a chance for W to be non-zero.

Let $\mathbb{W}$ be the vector space of possible W, and let p be a probability distribution over it. Assume further that p is symmetric -- that for all $W \in \mathbb{W}$, $p(W) = p(-W)$.

Then the pernicious effect of Goodhart's law is in full effect: suppose we ask an agent to maximise $U = V + W$, with V known and p giving the distribution over possible W.

Then, given that uncertainty, it will choose the policy π that maximises

- $E(U \mid \pi) = E(V + \sum_{W \in \mathbb{W}} p(W) \mid \pi) = E(V) + 0$,

since W and $-W$ cancel out.

So the agent will blindly maximise the proxy V.

We know: maximising behaviour is bad

But, most of the time, when we talk about Goodhart's law, we don't just mean "a proxy exists". We mean that not only does a proxy exist, but that maximising the proxy too much is pernicious for the true utility.

Consider for example a nail factory, where U is the number of true nails produced, and V is the number of "straight pieces of metal" produced. Here $W = U - V$ is the difference between the number of true nails and the pieces of metal.

In this case, we expect a powerful agent maximising V to do much, much worse on the U scale. As the agent expands and gets more control over the world's metal production, V continues to climb, while U tumbles. So V is not only a (bad) proxy for U; at the extremes, its pernicious.

But what if we considered $U' = V - W$? This is an odd utility indeed; it's equal to $V - (U - V) = 2V - U$. This is twice the number of pieces of metal produced, *minus* the number of true nails produced. And as the agent's power increases, so does V, but so does U', to an even greater extent. Now, U' won't increase to the same extent as it would under the optimal policy for U', but it still increases massively under V-optimisation.

And, implicitly, when we talk about Goodhart's law, we generally mean that the true utility is of type U, rather than of type U'; indeed, that things like U' don't really make sense as candidates for the true utility. So there is a break in symmetry between W and $-W$.

Putting this knowledge into figures

So, suppose π_0 is some default policy, and π_V^* is the V-maximising policy. One way of phrasing the stronger type of Goodhart's law is that:

- $E(U \mid \pi_0) > E(U \mid \pi_V^*)$.

Then, rearranging these gives:

- $E(W \mid \pi_0) - E(W \mid \pi_V^*) > E(V \mid \pi_V^*) - E(V \mid \pi_0) = C$.

Because π_V^* is the optimal policy for V, the term C is non-negative (and most likely strictly positive). So the new restriction on W is that:

- $E(W \mid \pi_0) - E(W \mid \pi_V^*) > C \geq 0$.

This is an affine restriction, in that if W and W' satisfy that restriction, then so does a mix $qW + (1-q)W'$ for $0 \leq q \leq 1$. In fact, it defines a hyperplane in W, with everything on one side satisfying that restriction, and everything on the other side not satisfying it.

In fact, the set that satisfies the restriction (call it W_+) is "smaller" (under p) than the set that does not (call that W_-). This is because, if $W \in W_+$, then $-W \in W_-$ -- but the converse is not true if $C > 0$.

And now, when an agent maximises $U = V + W$, with known V and W distributed by p but also known to obey that restriction, the picture is very different. It will maximise $V + W'$, where $W' = \sum_{W \in W_+} /p(W_+)$, and is far from 0 in general. So the agent won't just maximise the proxy.

Conclusion

So, even the seemingly trivial fact that we expect a particular type of Goodhart effect - even that trivial fact dramatically reduces the effect of Goodhart's law.

Now, the effect isn't enough to converge on a good U: we'll need to use other information for that. But note one interesting point: the more powerful the agent is, the more effective it is at maximising V, so the higher $E(V \mid \pi_V^*)$ gets -- and thus the higher C becomes. So the most powerful agents have the strongest restrictions on what the possible U's are. Note that we might be able to get this effect even for more

limited agents, by defining π^*_V not only as the optimal policy, but as some miraculous optimal policy where things work out unexpectedly well for the agent.

It will be interesting to see what happens as in situations where we account for more and more of our (implicit and explicit) knowledge about U.

On Collusion - Vitalik Buterin

This is a linkpost for https://hackernoon.com/on-collusion-fk1xr3png

> ...if there is a situation with some fixed pool of resources and some currently established mechanism for distributing those resources, and it's unavoidably possible for 51% of the participants can conspire to seize control of the resources, no matter what the current configuration is there is always some conspiracy that can emerge that would be profitable for the participants.
>
> **...This fact, the instability of majority games under cooperative game theory, is arguably highly underrated as a simplified general mathematical model of why there may well be no "end of history" in politics and no system that proves fully satisfactory; I personally believe it's much more useful than the more famous Arrow's theorem, for example.**

I've found this post quite useful in thinking about mechanism design and problems of designing stable systems.

The problem/solution matrix: Calculating the probability of AI safety "on the back of an envelope"

> If It’s Worth Doing, It’s Worth Doing With Made-Up Statistics
>
> -- Scott Alexander

My earlier post Three Stories for How AGI Comes Before FAI, was an "AI research strategy" post. "AI research strategy" is not the same as "AI strategy research". "AI research strategy" relates to the Hamming question for technical AI alignment researchers: What are the most important questions in technical AI alignment and why aren't you trying to answer them?

Here's a different way of looking at these questions (which hopefully has a different set of flaws).

Suppose we had a matrix where each column corresponds to an AI safety problem and each row corresponds to a proposed safety measure. Each cell contains an estimate of the probability of that safety measure successfully addressing that problem. If we assume measure successes are independent, we could estimate the probability that any given problem gets solved if we build an AGI which applies all known safety measures. If we assume problem successes are independent, we could estimate the probability that all known problems will be solved.

What about unknown problems? Suppose that God has a list of all AI safety problems. Suppose every time an AI alignment researcher goes out to lunch, there's some probability that they'll think of an AI safety problem chosen uniformly at random from God's list. In that case, if we know the number of AI alignment researchers who go out to lunch on any given day, and we also know the date on which any given AI safety problem was first discovered, then this is essentially a card collection process which lets us estimate the length of God's list as an unknown parameter (say, using maximum likelihood on the intervals between discovery times, or you could construct a prior based on the number of distinct safety problems in other engineering fields). In lay terms, if no one has proposed any new AI safety problems recently, it's plausible there aren't that many problems left.

How to estimate the probability that unknown problems will be successfully dealt with? Suppose known problems are representative of the entire distribution of problems. Then we can estimate the probability that an unknown problem will be successfully dealt with as follows: For each known problem, cover up rows corresponding to the safety measures inspired by that problem, then compute its new probability of success. That gives you a distribution of success probabilities for "unknown" problems. Assume success probabilities for actually-unknown problems are sampled from that distribution. Then you could compute the probability of any given actually-unknown probability being solved by computing the expected value of that distribution. Which also lets you compute the probability that your AGI will solve all the unknown problems too (assuming problem probabilities are independent).

What could this be useful for? Doing this kind of analysis could help us know whether it's more valuable to discover more problems or discover more safety measures on the current margin. It could tell us which problems are undersolved and would benefit from more people attacking them. Even without doing the analysis, you can see that trying to solve multiple problems with the same measure could be a good idea, since those measures are more likely to generalize to unknown problems. If overall success odds aren't looking good, we could make our first AI some kind of heavily restricted tool AI which tries to augment the matrix with additional rows and columns. If success odds *are* looking good, we could compare success odds with background odds of x-risk and try to figure out whether to actually turn this thing on.

Obviously there are many simplifying assumptions being made with this kind of "napkin math". For example, it could be that the implementation of safety measure A interacts negatively with the implementation of safety measure B and reverses its effect. It could be that we aren't sampling from God's list uniformly at random and some problems are harder to think of than others. Whether this project is actually worth doing is an "AI research strategy strategy" question, and thus above my pay grade. If it's possible to generate the matrix automatically using natural language processing on a corpus including e.g. the AI Alignment Forum, I guess that makes the project look more attractive.

What's going on with "provability"?

Every so often I hear seemingly mathematical statements involving the concept of being *provable*. For example:

- I've seen Gödel's Incompleteness Theorem stated as "if a mathematical system is powerful enough to express arithmetic, then either it contains a contradiction or there are true statements that it cannot prove."
- On the AI alignment forum, one of the pinned sequences describes Löb's Theorem as "If Peano Arithmetic can prove that a proof of P would imply the truth of P, then it can also prove P is true".

I find each of these statements baffling for a different reason:

- Gödel: What could it mean for a statement to be "true but not provable"? Is this just because there are some statements such that neither P nor not-P can be proven, yet one of them must be true? If so, I would (stubbornly) contest that perhaps P and not-P really are both non-true.
- Löb: How can a system of *arithmetic* prove anything? Much less prove things *about proofs*?

And I also have one more general confusion. What systems of reasoning could these kinds of theorems be set in? For example, I've heard that there are proofs that PA is consistent. Let's say one of those proofs is set in Proof System X. Now how do we know that Proof System X is consistent? Perhaps it can be proven consistent by using Proof System Y? Do we just end up making an infinite chain of appeals up along a tower of proof systems? Or do we eventually drive ourselves into the ground by reaching system that *nobody* could deny is valid? If so, why don't we just stop and PA or ZFC?

Oh, speaking of ZFC. There seems to be a debate about whether we should accept the Axiom of Choice. Isn't it...obviously true? I don't really understand this topic well enough to have any particular question about the AC debate, but my confusion definitely extends to that region of concept space.

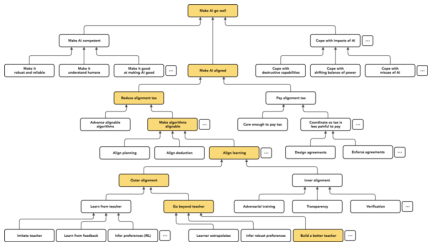
So here's my question: **Where can I learn about "provability" and/or what clarifying insights could you share about it?**

AI alignment landscape

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

Here (link) is a talk I gave at EA Global 2019, where I describe how intent alignment fits into the broader landscape of “making AI go well,” and how my work fits into intent alignment. This is particularly helpful if you want to understand what I’m doing, but may also be useful more broadly. I often find myself wishing people were clearer about some of these distinctions.

Here is the main overview slide from the talk:



The highlighted boxes are where I spend most of my time.

Here are the full slides from the talk.

Reasons for Hope & Objection Preemption (Novum Organum Book 1: 108-130)

This is the eighth post in the Novum Organum sequence. For context, see the sequence introduction.

We have used Francis Bacon's Novum Organum in the version presented at www.earlymoderntexts.com. Translated by and copyright to Jonathan Bennett. Prepared for LessWrong by Ruby.

Ruby's Reading Guide

> Novum Organum is organized as two books each containing numbered "aphorisms." These vary in length from three lines to sixteen pages. Bracketed titles of posts in this sequence, e.g. *Idols of the Mind Pt. 1*, are my own and do not appear in the original.
>
> While the translator, Bennett, encloses his editorial remarks in a single pair of [brackets], I have enclosed mine in a [[double pair of brackets]].

Bennett's Reading Guide

> [Brackets] enclose editorial explanations. Small ·dots· enclose material that has been added, but can be read as though it were part of the original text. Occasional •bullets, and also indenting of passages that are not quotations, are meant as aids to grasping the structure of a sentence or a thought. Every four-point ellipsis . . . . indicates the omission of a brief passage that seems to present more difficulty than it is worth. Longer omissions are reported between brackets in normal-sized type.

Aphorism Concerning the Interpretation of Nature: Book 1: 108–130

by Francis Bacon

[[Bacon continues his listing of reasons we should believe much greater scientific progress is possible.]]

108. That's all I have to say about getting rid of despair and creating **hope** by banishing or fixing past errors. Now, what other ways are there of creating **hope**? Here's a thought that occurs at once: Many useful discoveries have been made accidentally by men who weren't looking for *them* but were busy about other things; so no-one can doubt that if men seek for something and are busy about *it*, proceeding in an orderly and not a slapdash way, they will discover far more. Of course it can happen occasionally that someone accidentally stumbles on a result that he wouldn't have found if he had searched hard for it, but on the whole the opposite is the case—·things are discovered by methodical searching that couldn't have been found by accident·. So, far better things, and more of them, and at shorter intervals, are to be **hoped** for from •hard thinking, hard focussed work and concentration than from ••lucky· accidents, undisciplined whims and the like, which until now have been the main source of discoveries.

109. Here is another ground for **hope**: Discoveries have sometimes been made that would have been almost unthinkable in advance, and would have been written off as impossible. Men think about the *new* in terms of the *old*: to questions about what the •future holds they bring an imagination indoctrinated and coloured by the •past. This is a terrible way of forming opinions, because streams fed by nature's springs don't run along familiar channels.

Suppose that before gunpowder was invented someone described it in terms of its effects—'There is a new invention by means of which the strongest towers and walls can be demolished from a long way off'. That would no doubt have set men thinking about how to increase the power of *catapults* and *wheeled ramming devices*. The notion of a fiery blast suddenly and forcefully expanding and exploding would hardly have entered into any man's mind or imagination, because nothing closely analogous to that had ever been seen. Well, except perhaps in earthquakes and lightning, but *they* wouldn't have been seen as relevant because they are mighty works of *nature* which *men* couldn't imitate.

Or suppose that before the discovery of silk someone had said: 'They've discovered new a kind of thread for use in clothing and furniture-coverings; it is finer, softer, more beautiful and *stronger* than linen or wool.' Men would have begun to think of some silky kind of plant or of very fine hair of some animal or of the feathers and down of birds; they would *not* have thought of a web woven by a tiny worm in great quantities and renewing itself yearly. If anyone *had* said anything about a worm, he'd have been laughed at as dreaming of a new kind of cobweb! [Bacon then gives a third example: the magnet.] Yet these things and others like them lay concealed from men for centuries, and when they did come to light it wasn't through science or any technical skill but by accident and coincidence. As I have remarked, they were so *utterly*

different in kind from anything previously known that they couldn't possibly have been discovered through a preconceived notion of them.

So there are strong grounds for **hoping** that nature has concealed in its folds many wonderfully useful •things that aren't related to or parallel with anything that is now known, and lie right outside our imaginative reach. As the centuries roll on, •they too will doubtless come to light of their own accord in some roundabout way, as did gunpowder and the others; but by the method I am discussing they can be presented and anticipated speedily, suddenly and all at once.

110. Other discoveries prove that this can happen: splendid discoveries are lying at our feet, and we step over them without seeing them. The discoveries of

- gunpowder,
- silk,
- the magnet,
- sugar,
- paper,

or the like may seem to depend on certain properties of things of and nature—·properties that might have been hard to discover·. But there is nothing in *printing* that isn't wide open and almost easy. All that was needed was to see that

- although it is harder to arrange letter-types than to write by hand, the two procedures differ in that once the types have been arranged any number of impressions can be made from them, whereas hand-writing provides only a single copy,

and to see that

- ink can be so thickened so that it does its job but doesn't run, especially when the type faces upwards and the ink is rolled onto it from above.

It was merely because they didn't notice *these* ·obvious· facts that men went for so many ages without this most beautiful invention which is so useful in the spreading of knowledge.

But the human mind is such a mess when it comes to this business of discoveries that it first •distrusts and then •despises itself:

- before the discovery: it is not credible that any such thing can be found,
- afterwards: it is incredible that the world should have missed it for so long!

And this very thing entitles us to some **hope**, namely the hope that there is a great mass of discoveries still to be made—not just ones that will have to be dug out by techniques that we don't yet have, but also ones that may come to light through our transferring, ordering and applying things that we do know already, this being done with the help of the experimental approach that I call 'literate' [**101**].

111. Another ground of **hope** should be mentioned. Let men reflect on their infinite expenditure of intellect, time, and means on things of *far* less use and value ·than the discoveries I am talking about·. If even a small part of this were directed to sound and solid studies, there is no difficulty that couldn't be overcome. I mention this ·matter of the use of resources· because a collection of Natural and Experimental History, as I

envisage it and as it ought to be, is a great—as it were, a *royal*—work, and I freely admit that it will involve much labour and expense.

[It will appear in Book **2-11** that the 'collection' Bacon talks of is an orderly written account of phenomena, experiments and their results, not a physical museum.]

112. In the meantime, don't be put off by *how many* particulars there are; rather, let this give you **hope**. ·The fact is that you will be in worse trouble if you *don't* engage with them·; for the •particular phenomena of nature are a mere handful compared to the ·great multitudes of· •things that human ingenuity can fabricate if it cuts itself off from the clarifying effects of reality. And this road ·through the study of *real* events· soon leads to open ground, whereas the other—·the route through *invented* theories and thought-experiments·— leads to nothing but endless entanglement. Until now men haven't lingered long with •experience; they have brushed past it on their way to the ingenious •theorizings on which they have wasted unthinkable amounts of time. But if we had someone at hand who could answer our questions of the form 'What are the *facts* about this matter?', it wouldn't take many years for us to discover all causes and complete every science [the Latin literally means 'to discover all causes and sciences'].

113. Men may take some **hope**, I think, from my own example (I'm not boasting; just trying to be useful). If you are discouraged ·about the chances of progress in the sciences·, look at me!

- I am busier with affairs of state than any other man of my time,
- I lose a lot of time to ill-health, and
- in this ·scientific· work I am wholly a pioneer, not following in anyone's tracks and not getting advice from anyone.

And yet, ·despite these three sources of difficulty·, I think I have pushed things on a certain amount by sticking to the true road and submitting my mind to reality. Well, then, think what might be expected (now that I have pointed out the way) from men

- with plenty of free time,
- ·in good health·, and
- working together, on the basis of previous work ·by others·.

Unlike the work of sheerly *thinking up* hypotheses, proper scientific work *can* be done collaboratively; the best way is for men's efforts (especially in collecting experimental results) to be exerted separately and then brought together. Men will begin to know their strength only when they go this way—with one taking charge of one thing and another of another, instead of all doing all the same things.

114. Lastly, even if the breeze of **hope** that blows on us from that New Continent were fainter and less noticeable than it is, still we have to *try*—unless we prefer to have minds that are altogether abject! The loss that may come from •not trying is much greater than what may come from ·trying and· •not succeeding: by •not trying we throw away the chance of an immense good; by •not succeeding we only incur the loss of a little human labour. But from what I have said (and from some things that I *haven't* said) it seems to me that there is more than enough **hope** not only •to get a vigorous man to *try* but also to make a sober-minded and wise man *believe* ·that he will succeed·.

115. That completes what I wanted to say about getting rid of the pessimism that has been one of the most powerful factors delaying and hindering the progress of the

sciences. I have also finished with the signs and causes of errors, of sluggishness and of the prevailing ignorance. ·I've said more about this than you might think·, because the more subtle causes—the ones that aren't generally noticed or thought about—come under what I said about the 'idols' of the human mind.

And this should also bring to an end the part of my Great Fresh Start [see note in **31**] that is devoted to *rejection*, which I have carried out through three refutations:

(1) the refutation of innate human reason left to itself [see Preface];
(2) the refutation of demonstrations [see **44** and **69**];
(3) the refutation of the accepted philosophical doctrines [see **60–62**].

I refuted these in the ·only· way I *could* do so, namely through signs and the evidence of causes. I couldn't engage in any other kind of confutation because I differ from my opponents both on first principles and on rules of demonstration.

So now it is time to proceed to the actual techniques for interpreting nature and to the rules governing them—except that there is *still* something that has to be said first! In this first book of aphorisms my aim has been to prepare men's minds not just for •*understanding* what was to follow but for •*accepting* it; and now that I have •cleared up and washed down and levelled the floor of the mind, I have to •get the mind into a good attitude towards the things I am laying before it—to *look kindly* on them, as it were. ·This has to be worked for·, because anything new will be confronted by prejudgments ·against it·, not only ones created by old opinions but also ones created by false ideas about what the new thing is going to be. So I shall try to create sound and true opinions about what I am going to propose; but this is only a stop-gap expedient—a kind of security deposit—to serve until I can make the stuff itself thoroughly known.

116. First, then, don't think that I want to found a new sect in philosophy—like the ancient Greeks and like some moderns such as Telesio, Patrizzi or Severinus. For that's not what I am up to; and I really don't think that human welfare depends much on what abstract opinions anyone has about nature and its workings. No doubt many old theories of this sort can be revived and many new ones introduced, just as many theories of the heavens can be supposed that fit the phenomena well enough but differ from each other; but I'm not working on such useless speculative matters.

My purpose, rather, is to see whether I can't provide humanity's power and greatness with firmer foundations and greater scope. I have achieved some results—scattered through some special subjects—that I think to be far more true and certain and indeed more fruitful than any that have so far been used (I have collected them in the •fifth part of my Fresh Start); but I don't yet have a complete theory of everything to propound. It seems that the time hasn't come for that. I can't hope to live long enough to complete the •sixth part (which is to present science discovered through the proper interpretation of nature); but I'll be satisfied if in the middle parts I conduct myself soberly and usefully, sowing for future ages the seeds of a purer truth, and not shying away from the start of great things. [See note in **31**.]

117. Not being the founder of a sect, I am not handing out bribes or promises of particular works. You may indeed think that because I talk so much about 'works' ·or 'results'· and drag everything over to *that*, I should produce some myself as a down-payment. Well, I have already clearly said it many times, and am happy now to say it again: my project is not to get

works from works or
experiments from experiments (like the •empirics),

but rather to get

causes and axioms from works and experiments,

and then to get

new works and experiments from those causes and axioms (like the •legitimate interpreters of nature).

[An 'empiric' is someone who is interested in *what* works but not in *why* it works; especially a physician of that sort, as referred to by Locke when he speaks of 'swallowing down opinions as silly people do empirics' pills, without knowing what they are made of or how they will work'.]

If you look at

- my Tables of Discovery that ·will· constitute the fourth part of the Fresh Start, and
- the examples of particulars that I present in the second part, ·i.e. the present work·, and
- my observations on the history that I ·will· sketch in the third part,

you won't need any great intellectual skill to see indications and outlines of many fine results all through this material; but I openly admit that the natural history that I have so far acquired, from books and from my own investigations, is too skimpy, and not verified with enough accuracy, to serve the purposes of legitimate interpretation.

To anyone who is abler and better prepared ·than I am· for mechanical pursuits, and who is clever at getting results from experiment, I say: By all means go to work snipping off bits from my history and my tables and apply them to getting results—this could serve as *interest* until the *principal* is available. But I am hunting for bigger game, and I condemn all hasty and premature interruptions for such things as these, which are (as I often say) like Atalanta's spheres. I don't go dashing off after golden apples, like a child; I bet everything on art's winning its race against nature. [On Atalanta and the race see **70**.] I don't scurry around clearing out moss and weeds; I wait for the harvest when the crop is ripe.

118. When my history and Tables of Discovery are read, it will surely turn out that some things in the experiments themselves are not quite certain or perhaps even downright false, which may lead you to think that the foundations and principles on which my discoveries rest are ·also· false and doubtful. But this doesn't matter, for such things are bound to happen at first. It's like a mere typographical error, which doesn't much hinder the reader because it is easy to correct as you read. In the same way, ·my· natural history may contain many experiments that are false, but it won't take long for them to be easily expunged and rejected through the discovery of causes and axioms. It is nevertheless true that if big mistakes come thick and fast in a natural history, they can't possibly be corrected or amended through any stroke of intelligence or skill. Now, my natural history has been collected and tested with great diligence, strictness and almost *religious* care, yet there may be errors of detail tucked away in it; so what should be said of run-of-the-mill natural history, which is so careless and *easy* in comparison with mine? And what of the philosophy and sciences

built on that kind of sand (or rather *quicksand*)? So no-one should be troubled by what I have said.

119. My history and experiments will contain many things that are

- trivial, familiar and ordinary, many that are
- mean and low [see **120**], and many that are
- extremely subtle, merely speculative, and seemingly useless [see **121**].

Such things could lead men to lose interest or to become hostile ·to what I have to offer. I shall give these one paragraph each·.

Men should bear in mind that until now *their* activities have consisted only in explaining unusual events in terms of more usual ones, and they have simply taken the usual ones for granted, not asking what explains *them*. So they haven't investigated the causes of

- weight,
- rotation of heavenly bodies,
- heat,
- cold,
- light,
- hardness,
- softness,
- rarity,
- density,
- liquidity,
- solidity,
- life,
- lifelessness,
- similarity,
- dissimilarity,
- organicness,

and the like. They have accepted these as self-evident and obvious, and have devoted their inquiring and quarrelling energies to less common and familiar things.

But I *have to* let the most ordinary things into my history, because I know that until we have properly looked for and found the causes of common things and the causes of those causes, we can't make judgments about uncommon or remarkable things, let alone bring anything new to light. Indeed, I don't think that anything holds up philosophy more than the fact that common and familiar events don't cause men to stop and think, but are received casually with no inquiry into their causes. A result of this we need •to pay attention to things that are known and familiar at least as often as •to get information about unknown things.

120. As for things that are low or even filthy: as Pliny says, these should be introduced with an apology, but they should be admitted into natural history just as the most splendid and costly things should. And that doesn't pollute the natural history that admits them; the sun enters the sewer as well as the palace, but isn't polluted by that! I am not building a monument dedicated to human glory or erecting a pyramid in its honour; what I'm doing is to lay a foundation for a holy temple in the human intellect—a temple modelled on the world. So I follow that model, because whatever is worthy of *being* is worthy of *scientific knowledge*, which is the image or likeness of being; and low things *exist* just as splendid ones do. And another point:

just as from certain putrid substances such as musk and civet the sweetest odours are sometimes generated, so also mean and sordid events sometimes give off excellent and informative light. That is enough about this; *more* than enough, because this sort of squeamishness is downright childish and effeminate.

121. The third objection must be looked into much more carefully. I mean the objection that many things in my history will strike ordinary folk, and indeed ·non-ordinary· ones trained in the presently accepted systems, as intricately subtle and *useless*. It is especially because of this objection that I have said, and should ·again· say, that in the initial stages ·of the inquiry· I am aiming at experiments of light, not experiments of fruit [see **99**]. In this, as I have often said [see **70**], I am following the example of the divine creation which on the first day produced nothing but light, and gave that a day to itself without doing any work with matter. To suppose, therefore, that things like these ·'subtleties' of mine· are useless is the same as supposing that light is useless because it isn't a *thing*, isn't solid or material. And well-considered and well-delimited knowledge of simple natures is like light: it gives entrance to all the secrets of nature's workshop, and has the power to gather up and draw after it whole squadrons of works and floods of the finest axioms; yet there is hardly anything we can *do with it* just in itself. Similarly the •letters of the alphabet taken separately are useless and meaningless, yet they're the basic materials for the planning and composition of all discourse. So again the •seeds of things have much latent power, but nothing comes of it except in their development. And ·light is like scientific subtleties in another way, namely·: the scattered rays of light don't do any good unless they are made to converge.

If you object to speculative subtleties, what will you say about the schoolmen [= 'mediaeval and early modern Aristotelians'], who have wallowed in subtleties? And *their* subtleties were squandered on •words (or on popular notions—same thing!) rather than on •facts or nature; and they were useless the whole way through, unlike mine, which are indeed useless right now but which promise endless benefits later on. But this is sure, and you should know it:

> All subtlety in disputations and other mental bustling about, if it occurs after the axioms have been discovered, comes too late and has things backwards. The true and proper time for subtlety, or anyway the chief time for it, is when pondering experiments and basing axioms on them.

For that other ·later· subtlety grasps and snatches at [*captat*] nature but can never get a grip on [*capit*] it. . . .

A final remark about the lofty dismissal from natural history of everything •common, everything •low, everything •subtle and as it stands useless: When a haughty monarch rejected a poor woman's petition as unworthy thing and beneath his dignity, she said: 'Then leave off being king.' That may be taken as an oracle. For someone who won't attend to things like •these because they are too paltry and minute can't take possession of the kingdom of nature and can't govern it.

122. This may occur to you: 'It is amazing that you have the nerve to push aside all the sciences and all the authorities at a single blow, doing this single-handed, without bringing in anything from the ancients to help you in your battle and to guard your flanks.'

Well, I know that if I had been willing to be so dishonest, I could easily have found support and honour for my ideas by referring them either •to ancient times before the

time of the Greeks (when natural science may have flourished more ·than it did later·, though *quietly* because it hadn't yet been run through the pipes and trumpets of the Greeks), or even, in part at least, •to some of the Greeks themselves. This would be like the men of no family who forge genealogical tables that 'show' them to come from a long line of nobility. But I am relying on the evidentness of ·the truth about· things, and I'll have nothing to do with any form of fiction or fakery. Anyway, it *doesn't matter* for the business in hand whether the discoveries being made now •were known to the ancients long ago and •have alternately flourished and withered through the centuries because of the accidents of history (just as it doesn't matter to mankind whether the New World is the island of Atlantis that the ancients knew about or rather is now discovered for the first time). It doesn't matter because *discoveries*—·even if they are *re*discoveries·—have to be sought [*petenda*] from the light of nature, not called back [*repetenda*] from the shadows of antiquity.

As for the fact that I am finding fault with everyone and everything: when you think about it you'll see that *that* kind of censure is more likely to be right than a partial one would be—and less damaging, too. For a partial censure would imply that the errors were not rooted in primary notions, and that there had been some true discoveries; they could have been used to correct the false results, ·and the people concerned would have been to blame for not seeing this·. But in fact the errors were fundamental; they came not so much from false judgment as from not attending to things that should be attended to; so it's no wonder that men haven't obtained what they haven't tried for, haven't reached a mark that they never set up, haven't come to the end of a road that they never started on.

As for the insolence that ·you might think· is inherent in what I am doing: if a man says that

> •his steady hand and good eyes enable him to draw a straighter line or a more perfect circle than anyone else,

he is certainly •making a comparison of abilities; but if he says only that

> •with the help of a ruler or a pair of compasses can draw a straighter line or a more perfect circle than anyone else can by eye and hand alone,

he isn't •making any great boast. And I'm saying this not only about these first initiating efforts of mine but also about everyone who tackles these matters in the future. For my route to discovery in the sciences puts men on the same intellectual level, leaving little to individual excellence, because it does everything by the surest rules and demonstrations. So I attribute my part in all this, as I have often said, to good luck rather than to ability—it's a product of *time* rather than of *intelligence*. For there's no doubt that luck has something to do with men's thoughts as well as with their works and deeds.

123. Someone once said jokingly 'It can't be that we think alike, when one drinks water and the other drinks wine'; and this nicely fits my present situation. Other men, in ancient as well as in modern times, have done their science drinking a crude liquor —like water

> **(1)** flowing spontaneously from a spring or **(2)** hauled up by wheels from a well,
> **(1)**flowing spontaneously from the intellect or **(2)** hauled up by logic.

Whereas I drink a toast with a liquor strained from countless grapes, ripe and fully seasoned ones that have been gathered and picked in clusters, squeezed in the press,

and finally purified and clarified in the vat. No wonder I am at odds with the others!

124. This also may occur to you: 'You say it against others, but it can be said against you, that the goal and mark that you have set up for the sciences is not the true or the best.' ·The accusation would develop like this·:

> Contemplation of *the truth* is a worthier and loftier thing than thinking about how *big and useful* one's practical results will be. Lingering long and anxiously on •experience and •matter and •the buzz of individual events drags the mind down to earth, or rather sinks it to an underworld of turmoil and confusion, dragging it away from a much more heavenly condition—the serene tranquillity of abstract wisdom.

Now I *agree* with this line of thought; what the objectors here point to as preferable is what I too am after, above everything else. For I am laying down in the human intellect the foundations for a *true model of the world*—the world as it turns out to be, not as one's reason would like it to be. This can't be done unless the world is subjected to a very diligent dissection and anatomical study. As for the stupid models of the world that men have dreamed up in philosophical systems—like the work of *apes*!—they should be utterly scattered to the winds. You need to know what a big difference there is (as I said above [**23**]) between the •idols of the human mind and the •ideas in the divine mind. The former are merely arbitrary abstractions; the latter are the creator's little seals on the things he has created, stamped into matter in true and exquisite lines. In these matters, therefore, truth and usefulness are the very same thing; and practical applications ·of scientific results· are of greater value as pledges of truth than as contributing to the comforts of life.

125. Or you may want to say this: 'You are only doing what the ancients did before you; so that you are likely, after all this grinding and shoving, to end up with one of the systems that prevailed in ancient times.' The case for this goes as follows:

> The ancients also provided at the outset of their speculations a great store and abundance of examples and particulars, sorted out and labelled in notebooks; then out of them they constructed their systems and techniques; and when after that they had checked out everything they published their results to the world with a scattering of examples for proof and illustration; but they saw no need to take the considerable trouble of publishing their working notes and details of experiments. So they did what builders do: after the house was built they removed the scaffolding and ladders out of sight.

I'm sure they did! But this objection (or misgiving, rather) will be easily answered by anyone who hasn't completely forgotten what I have said above. The form of inquiry and discovery that the ancients used—they declared it openly, and it appears on the very face of their writings—was simply this:

> From a few examples and particulars (with some common notions thrown in, and perhaps some of the most popular accepted opinions). they rushed to the most general conclusions, the ·would-be· first principles of ·their· science. Taking the truth of *these* as fixed and immovable, they proceeded to derive from them—through intermediate propositions— lower-level conclusions out of which they built their system. Then if any new particulars and examples turned up that didn't fit their views, they either •subtly moulded them into their system by distinctions or explanations of their rules, or •coarsely got rid of them by ·tacking· exceptions ·onto their principles·. As for particulars that weren't in conflict ·with their views·,

> they laboured away through thick and thin to assign them causes in conformity with their principles.

But this wasn't the experimental natural history that was wanted; far from it. And anyway dashing off to the highest generalities ruined everything.

126. will occur to you too: 'By forbidding men to announce principles and take them as established until they have arrived at the highest generalities in the right way through intermediate steps, you are inviting them to *suspend judgment*, bringing this whole affair down to Acatalepsy.' Not so. What I have in mind and am propounding is not Acatalepsy [from Greek, = 'the doctrine that nothing can be understood'] but rather Eucatalepsy [from Greek, = 'the provision of what is needed for things to be understood']. I don't •disparage the senses, I •serve them; I don't •ignore the intellect, I •regulate it. And it is surely better that we should

> know everything that we need to know, while thinking that our knowledge doesn't get to the heart of things

than that we should

> think our knowledge gets to the heart of things, while we don't yet know anything we need to know.

127. You may want to ask—just as a query, not an objection—whether I am talking only about natural philosophy, or whether instead I mean that the other sciences—logic, ethics and politics—should be conducted in my way. Well, I certainly mean what I have said to apply to them all. Just as •common logic (which rules things by syllogisms) extends beyond natural sciences to all sciences, so does •mine (which proceeds by induction) also embrace everything. I am constructing a history and table of discovery for

> •anger, fear, shame, and the like; for
> •matters political; and for
> •the mental operations of memory, composition and division, judgment and the rest,

just as much as for

> •heat and cold, light, vegetative growth and the like.

But my method of interpretation ·differs from the common logic in one important respect; my method·, after the history has been prepared and set in order, concerns itself not only with •the movements and activities of the mind (as the common logic does) but also with •the nature of things ·outside the mind·. I guide the mind so that its way of engaging with any particular thing is always appropriate. That's why my doctrine of interpretation contains many different instructions, fitting the discovery-method according to the quality and condition of the subject-matter of the inquiry.

128. 'Do you want to pull down and destroy the philosophy, arts and sciences that are now practised?' There ought to be no question about that. Far from wanting to destroy them, I am very willing to see them used, developed and honoured. I don't want to get in the way of their •giving men something to dispute about, •supplying decoration for discourse, •providing the 'experts' with an income, and •facilitating civil life—acting, in short, like coins that have value because men agree to give it to them. Let me clear about this: what I am presenting won't be much use for purposes such as those, since

it can't be brought within reach of the minds of the vulgar except ·indirectly·, through effects and works. My published writings, especially my *Two Books on the Advancement of Learning*, show well enough the sincerity of my declaration of friendly good will toward the accepted sciences, so I shan't expend more words on that topic here. Meanwhile I give clear and constant warning that the methods now in use won't lead to any great progress in the theoretical parts of the sciences, and won't produce much in the way of applied-science results either.

129. All that remains for me to say are a few words about the excellence of the end in view. If I had said them earlier they might have seemed like mere *prayers*; but perhaps they'll have greater weight now, when hopes have been created and unfair prejudices removed. I wouldn't have said them even now if I had done the whole job myself, not calling on anyone else to help with the work, because ·words said in praise of the object of this exercise· might be taken as a proclamation of my own deserts. But ·I'm not going it alone·; I do want to energize others and kindle their zeal, so it is appropriate that I put men in mind of some things, ·even at the risk of *seeming* to boast·.

The making of great ·scientific· discoveries seems to have pride of place among human actions. That was the attitude of the ancients: they honoured the makers of discoveries as though they were *gods*, but didn't go higher than *demigods* in their honours for those who did good service in the state (founders of cities and empires, legislators, saviours of their country from long endured evils, quellers of tyrannies, and the like). And if you think accurately about the two ·kinds of benefactor· you will see that the ancients were right about them. Why? **(1)** Because the benefits of ·scientific· discoveries can •extend to the whole of mankind, and can •last for all time, whereas civil benefits •apply only to particular places and •don't last for very long.

(2) Also, improvements in civil matters usually bring violence and confusion with them, whereas ·scientific· discoveries bring delight, and confer benefits without causing harm or sorrow to anyone.

·Scientific· discoveries are like new creations, imitations of God's works. . . . It seems to be worth noting that Solomon, the marvel of the world, though mighty in empire and in gold, in the magnificence of his works, his court, his household, his fleet, and the lustre of his name, didn't glory in any of these, but pronounced that 'It is the glory of God to conceal a thing; but the honour of kings is to search out a matter' (*Proverbs* 25:2).

If you compare how men live in the most civilized provinces of Europe with how they live in the wildest and most barbarous areas of the American continent, you will think the difference is big enough—the difference in •the condition of the people in themselves as well as in •what conveniences and comforts they have available to them—to justify the saying that 'man is a god to man'. And this difference doesn't come from the Europeans' having better soil, a better climate, or better physiques, but from the arts [see note on 'art' here].

Notice the *vigour* of discoveries, their power to generate consequences. This is nowhere more obvious than in three discoveries that the ancients didn't know and whose origins (all quite recent) were obscure and humdrum. I am talking about the arts of •printing, •gunpowder, and •the nautical compass. These three have changed the whole aspect and state of things throughout the world—the first in literature, the second in warfare, the third in navigation—bringing about countless changes; so that there seems to have been no empire, no philosophical system, no star that has

exerted greater power and influence in human affairs than these mechanical discoveries.

For my next point, I need to distinguish the three kinds— three *levels*, as it were—of human ambition. **(1)** Some people want to extend their power within their own country, which is a commonplace and inferior kind of ambition. **(2)** Some work to extend the power and dominion of their country in relation to mankind in general; this is certainly not as base as **(1)** is, but it is just as much a case of greed. **(3)** If a man tries to get mankind's power and control over the universe off to a fresh start, and to extend it, *his*ambition (if it is ambition at all) is certainly more wholesome and noble ·than the other two·. Now—·this being the point I wanted to make·—man's control over things depends wholly on the arts and sciences, for we can't command nature except by obeying her.

A further point: it sometimes happens that •one particular discovery is so useful to mankind that the person who made it and thus put the whole human race into his debt is regarded as superhuman; so how much higher a thing it is to discover something through which •everything else can easily be discovered! ·Not that a discovery's consequences are the main thing about it·. *Light* is useful in countless ways, enabling us to walk, practise our arts, read, recognize one another, and yet something that is finer and lovelier than all those uses of light is *seeing light*. Similarly, merely contemplating things as they are, without superstition or imposture, error or confusion, is in itself worthier than all the practical upshots of discoveries.

Final point: If anyone counts it against the arts and sciences that they can be debased for purposes of wickedness, luxury, and the like, don't be influenced by that. The same can be said of all earthly goods: intelligence, courage, strength, beauty, wealth —even *light*! Just let the human race get back the right over nature that God gave to it, and give it scope; how it is put into practice will be governed by sound reason and true religion.

130. The time has come for me to present the art of interpreting nature—the art itself, ·not just remarks about the need for it, its virtues, and so on·. Although I think I have given true and most useful precepts in it, I don't say that this art is absolutely necessary, implying that nothing could be done without it. In fact, I think that if

> •men had ready at hand a sound history of nature and of experiments, •were thoroughly practised in it, and •imposed on themselves two rules: **(1)** set aside generally accepted opinions and notions, and **(2)** for a while keep your mind away from the highest and second-to-highest generalizations,

they would arrive at my form of interpretation sheerly through their own natural intelligence, with no help from any other rules or techniques. For interpretation is the true and natural work of the mind when it is freed from blockages. It is true, however, that it can all be done more readily and securely with help from my precepts.

And I don't say, either, that my art of interpreting nature is complete so that nothing can be added to it. On the contrary: I am concerned with the mind not only in respect of its own capacities but also in respect of how it engages with things; so I have to think that the art of discovery can develop as more discoveries are made.

The next post in the sequence will be posted Thursday, October 24 at latest by 4:00pm PDT.

Human-AI Collaboration

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://bair.berkeley.edu/blog/2019/10/21/coordination/

We've just released our paper on human-AI collaboration. The paper makes a straightforward-to-me point that self-play training is not going to work as well with humans in collaborative settings as in competitive settings. Basically, humans cause a distributional shift for the self-play agent. However, in the competitive case, the self-play agent should move towards the minimax policy, which has the nice property of guaranteeing a certain level of reward regardless of the opponent. The collaborative case has no such guarantee, and the distribution shift can tank the team performance. We demonstrated this empirically on a simplified version of the couch coop game Overcooked (which is amazing, I've played through both Overcooked games with friends).

As with a previous post, the rest of this post assumes that you've already read the blog post. I'll speculate about how the general area of human-AI collaboration is relevant for AI alignment. Think of these as rationalizations of the research after the fact.

It's necessary for assistance games

Assistance games (formerly called CIRL games) involve a human and an agent working together to optimize a shared objective that only the human knows. I think the general framework makes a lot of sense. Unfortunately, assistance games are extremely intractable to solve. If you try to scale up assistance games as a whole, the resulting environment is not very strategically complex, because it's hard to do preference learning and coordination simultaneously with deep RL. This suggests trying to make progress on subproblems within assistance games.

Usually, when people talk about making progress on "the CIRL agenda", they are talking about the preference learning aspect of an assistance game. We typically simplify to a single-agent setting and do preference learning, as in learning from comparisons or demonstrations. However, a useful agent will also need to properly coordinate with the human in order to be efficient. This suggests work on human-AI collaboration. We can work on this problem independently of preference learning simply by assuming that the agent knows the true reward function. This is exactly the setting that we study.

In general, I expect that if one hopes to take an assistance-game-like approach to AI alignment, work on human-AI collaboration will be necessary. The main uncertainty is whether assistance games are the right approach. Under a learning-based model of AI development, I think it is reasonably likely that the assistance game paradigm will be useful, without solving all problems (in particular, it may not solve inner alignment).

It seems important to figure out coordination

Regardless of whether we use assistance games, it's probably worthwhile to figure out how an AI system should coordinate with another agent that is not like itself. I don't

have a concrete story here; it’s just a general broad intuition.

It leads to more human-AI research

On my model, the best reason for optimism is that researchers will try to build useful AI systems, they’ll run into problems, and then they’ll fix those problems. Under this model, a useful intervention to run is to discover the problems sooner. This isn’t completely clear, since maybe if you discover the problems sooner, the root causes aren’t as obvious, and you are less likely to fix the entire problem -- but I think the main effect is in fact an increase in safety.

This would be my guess for how this research will most impact AI safety. We (by which I mean mostly Micah and somewhat me) spent a bunch of time cleaning up the code, making it easy for others to work with, creating nice figures, writing up a good blog post, etc. in an effort to have other ML researchers actually make progress on these issues. (However, I wouldn’t be too surprised if other researchers used the environment, but for a different purpose.)

[AN #69] Stuart Russell's new book on why we need to replace the standard model of AI

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

This is a bonus newsletter summarizing Stuart Russell's new book, along with summaries of a few of the most relevant papers. It's entirely written by Rohin, so the usual "summarized by" tags have been removed.

We're also changing the publishing schedule: so far, we've aimed to send a newsletter every Monday; we're now aiming to send a newsletter every Wednesday.

Audio version here (may not be up yet).

Human Compatible: Artificial Intelligence and the Problem of Control *(Stuart Russell)*: *Since I am aiming this summary for people who are already familiar with AI safety, my summary is substantially reorganized from the book, and skips large portions of the book that I expect will be less useful for this audience. If you are not familiar with AI safety,* ***note that I am skipping many arguments and counterarguments in the book that are aimed for you****. I'll refer to the book as "HC" in this newsletter.*

Before we get into details of impacts and solutions to the problem of AI safety, it's important to have a model of how AI development will happen. Many estimates have been made by figuring out the amount of compute needed to run a human brain, and figuring out how long it will be until we get there. HC doesn't agree with these; it suggests the bottleneck for AI is in the algorithms rather than the hardware. We will need several conceptual breakthroughs, for example in language or common sense understanding, cumulative learning (the analog of cultural accumulation for humans), discovering hierarchy, and managing mental activity (that is, the metacognition needed to prioritize what to think about next). It's not clear how long these will take, and whether there will need to be more breakthroughs after these occur, but these seem like necessary ones.

What could happen if we do get beneficial superintelligent AI? While there is a lot of sci-fi speculation that we could do here, as a weak lower bound, it should at least be able to automate away almost all existing human labor. Assuming that superintelligent AI is very cheap, most services and many goods would become extremely cheap. Even many primary products such as food and natural resources would become cheaper, as human labor is still a significant fraction of their production cost. If we assume that this could bring up everyone's standard of life up to that of the 88th percentile American, that would result in nearly a *tenfold* increase in world GDP per year. Assuming a 5% discount rate per year, this corresponds to $13.5 *quadrillion*

net present value. Such a giant prize removes many reasons for conflict, and should encourage everyone to cooperate to ensure we all get to keep this prize.

Of course, this doesn't mean that there aren't any problems, even with AI that does what its owner wants. Depending on who has access to powerful AI systems, we could see a rise in automated surveillance, lethal autonomous weapons, automated blackmail, fake news and behavior manipulation. Another issue that could come up is that once AI is better than humans at all tasks, we may end up delegating everything to AI, and lose autonomy, leading to *human enfeeblement*.

This all assumes that we are able to control AI. However, we should be cautious about such an endeavor -- if nothing else, we should be careful about creating entities that are more intelligent than us. After all, the gorillas probably aren't too happy about the fact that their habitat, happiness, and existence depends on our moods and whims. For this reason, HC calls this the *gorilla problem*: specifically, "the problem of whether humans can maintain their supremacy and autonomy in a world that includes machines with substantially greater intelligence". Of course, we aren't in the same position as the gorillas: we get to *design* the more intelligent "species". But we should probably have some good arguments explaining why our design isn't going to succumb to the gorilla problem. This is especially important in the case of a fast intelligence explosion, or *hard takeoff*, because in that scenario we do not get any time to react and solve any problems that arise.

Do we have such an argument right now? Not really, and in fact there's an argument that we *will* succumb to the gorilla problem. The vast majority of research in AI and related fields assumes that there is some definite, known *specification* or *objective* that must be optimized. In RL, we optimize the *reward function*; in search, we look for states matching a *goal criterion*; in statistics, we minimize *expected loss*; in control theory, we minimize the *cost function* (typically deviation from some desired behavior); in economics, we design mechanisms and policies to maximize the *utility* of individuals, *welfare* of groups, or *profit* of corporations. This leads HC to propose the following standard model of machine intelligence: *Machines are intelligent to the extent that their actions can be expected to achieve their objectives.* However, if we put in the wrong objective, the machine's obstinate pursuit of that objective would lead to outcomes we won't like.

Consider for example the content selection algorithms used by social media, typically maximizing some measure of engagement, like click-through. Despite their lack of intelligence, such algorithms end up changing the user's preference so that they become more predictable, since more predictable users can be given items they are more likely to click on. In practice, this means that users are pushed to become more extreme in their political views. Arguably, these algorithms have already caused much damage to the world.

So the problem is that we don't know how to put our objectives inside of the AI system so that when it optimizes its objective, the results are good for us. Stuart calls this the "King Midas" problem: as the legend goes, King Midas wished that everything he touched would turn to gold, not realizing that "everything" included his daughter and his food, a classic case of a badly specified objective (AN #1). In some sense, we've known about this problem for a long time, both from King Midas's tale, and in stories about genies, where the characters inevitably want to undo their wishes.

You might think that we could simply turn off the power to the AI, but that won't work, because for almost any definite goal, the AI has an incentive to stay operational, just

because that is necessary for it to achieve its goal. This is captured in what may be Stuart's most famous quote: *you can't fetch the coffee if you're dead*. This is one of a few worrisome convergent instrumental subgoals.

What went wrong? The problem was the way we evaluated machine intelligence, which doesn't take into account the fact that machines should be useful for *us*. HC proposes: *Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives*. But with this definition, instead of our AI systems optimizing a definite, wrong objective, they will *also* be uncertain about the objective, since we ourselves don't know what our objectives are. HC expands on this by proposing three principles for the design of AI systems, that I'll quote here in full:

1. *The machine's only objective is to maximize the realization of human preferences.*

2. *The machine is initially uncertain about what those preferences are.*

3. *The ultimate source of information about human preferences is human behavior.*

Cooperative Inverse Reinforcement Learning provides a formal model of an *assistance game* that showcases these principles. You might worry that an AI system that is uncertain about its objective will not be as useful as one that knows the objective, but actually this uncertainty is a feature, not a bug: it leads to AI systems that are deferential, that ask for clarifying information, and that try to learn human preferences. The Off-Switch Game shows that because the AI is uncertain about the reward, it will let itself be shut off. These papers are discussed later in this newsletter.

So that's the proposed solution. You might worry that the proposed solution is quite challenging: after all, it requires a shift in the entire way we do AI. What if the standard model of AI can deliver more results, even if just because more people work on it? Here, HC is optimistic: the big issue with the standard model is that it is not very good at learning our preferences, and there's a huge economic pressure to learn preferences. For example, I would pay a lot of money for an AI assistant that accurately learns my preferences for meeting times, and schedules them completely autonomously.

Another research challenge is how to actually put principle 3 into practice: it requires us to connect human behavior to human preferences. Inverse Reward Design and Preferences Implicit in the State of the World (AN #45) are example papers that tackle portions of this. However, there are *lots* of subtleties in this connection. We need to use *Gricean semantics* for language: when we say X, we do not mean the literal meaning of X: the agent must also take into account the fact that we bothered to say X, and that we didn't say Y. For example, I'm only going to ask for the agent to buy a cup of coffee if I believe that there is a place to buy reasonably priced coffee nearby. If those beliefs happen to be wrong, the agent should ask for clarification, rather than trudge hundreds of miles or pay hundreds of dollars to ensure I get my cup of coffee.

Another problem with inferring preferences from behavior is that humans are nearly always in some deeply nested plan, and many actions don't even occur to us. Right now I'm writing this summary, and not considering whether I should become a fireman. I'm not writing this summary because I just ran a calculation showing that this would best achieve my preferences, I'm doing it because it's a subpart of the overall plan of writing this bonus newsletter, which itself is a subpart of other plans. The connection to my preferences is very far up. How do we deal with that fact?

There are perhaps more fundamental challenges with the notion of "preferences" itself. For example, our *experiencing self* and our *remembering self* may have different preferences -- if so, which one should our agent optimize for? In addition, our preferences often change over time: should our agent optimize for our current preferences, even if it knows that they will predictably change in the future? This one could potentially be solved by learning *meta-preferences* that dictate what kinds of preference change processes are acceptable.

All of these issues suggest that we need work across many fields (such as AI, cognitive science, psychology, and neuroscience) to reverse-engineer human cognition, so that we can put principle 3 into action and create a model that shows how human behavior arises from human preferences.

So far, we've been talking about the case with a single human. But of course, there are going to be multiple humans: how do we deal with that? As a baseline, we could imagine that every human gets their own agent that optimizes for their preferences. However, this will differentially benefit people who care less about other people's welfare, since their agents have access to many potential plans that wouldn't be available to an agent for someone who cared about other people. For example, if Harriet was going to be late for a meeting with Ivan, her AI agent might arrange for Ivan to be even later.

What if we had laws that prevented AI systems from acting in such antisocial ways? It seems likely that superintelligent AI would be able to find loopholes in such laws, so that they do things that are strictly legal but still antisocial, e.g. line-cutting. (This problem is similar to the problem that we can't just write down what we want and have AI optimize it.)

What if we made our AI systems utilitarian (assuming we figured out some acceptable method of comparing utilities across people)? Then we get the "Somalia problem": agents will end up going to Somalia to help the worse-off people there, and so no one would ever buy such an agent.

Overall, it's not obvious how we deal with the transition from a single human to multiple humans. While HC focuses on a potential solution for the single human / single agent case, there is still much more to be said and done to account for the impact of AI on all of humanity. To quote HC, "There is really no analog in our present world to the relationship we will have with beneficial intelligent machines in the future. It remains to be seen how the endgame turns out."

Rohin's opinion: I enjoyed reading this book; I don't usually get to read a single person's overall high-level view on the state of AI, how it could have societal impact, the argument for AI risk, potential solutions, and the need for AI governance. It's nice to see all of these areas I think about tied together into a single coherent view. While I agree with much of the book, especially the conceptual switch from the standard model of intelligent machines to Stuart's model of beneficial machines, I'm going to focus on disagreements in this opinion.

First, the book has an implied stance towards the future of AI research that I don't agree with: I could imagine that powerful AI systems end up being created by learning alone without needing the conceptual breakthroughs that Stuart outlines. This has been proposed in e.g. AI-GAs (AN #63)), and seems to be the implicit belief that drives OpenAI and DeepMind's research agendas. This leads to differences in risk analysis and solutions: for example, the inner alignment problem (AN #58) only applies to

agents arising from learning algorithms, and I suspect would not apply to Stuart's view of AI progress.

The book also gives the impression that to solve AI safety, we simply need to make sure that AI systems are optimizing the right objective, at least in the case where there is a single human and a single robot. Again, depending on how future AI systems work, that could be true, but I expect there will be other problems that need to be solved as well. I've already mentioned inner alignment; other graduate students at CHAI work on e.g. robustness and transparency.

The proposal for aligning AI requires us to build a model that relates human preferences to human behavior. This sounds extremely hard to get completely right. Of course, we may not need a model that is completely right: since reward uncertainty makes the agent amenable to shutdowns, it seems plausible that we can correct mistakes in the model as they come up. But it's not obvious to me that this is sufficient.

The sections on multiple humans are much more speculative and I have more disagreements there, but I expect that is simply because we haven't done enough research yet. For example, HC worries that we won't be able to use laws to prevent AIs from doing technically legal but still antisocial things for the benefit of a single human. This seems true if you imagine that a single human suddenly gets access to a superintelligent AI, but when everyone has a superintelligent AI, then the current system where humans socially penalize each other for norm violations may scale up naturally. The overall effect depends on whether AI makes it easier to violate norms, or to detect and punish norm violations.

Read more: Max Tegmark's summary, Alex Turner's thoughts

AI Alignment Podcast: Human Compatible: Artificial Intelligence and the Problem of Control *(Lucas Perry and Stuart Russell)*: This podcast covers some of the main ideas from the book, which I'll ignore for this summary. It also talks a bit about the motivations for the book. Stuart has three audiences in mind. He wants to explain to laypeople what AI is and why it matters. He wants to convince AI researchers that they should be working in this new model of beneficial AI that optimizes for our objectives, rather than the standard model of intelligent AI that optimizes for its objectives. Finally, he wants to recruit academics in other fields to help connect human behavior to human preferences (principle 3), as well as to figure out how to deal with multiple humans.

Stuart also points out that his book has two main differences from Superintelligence and Life 3.0: first, his book explains how existing AI techniques work (and in particular it explains the standard model), and second, it proposes a technical solution to the problem (the three principles).

Cooperative Inverse Reinforcement Learning *(Dylan Hadfield-Menell et al)*: This paper provides a formalization of the three principles from the book, in the case where there is a single human H and a single robot R. H and R are trying to optimize the same reward function. Since both H and R are represented in the environment, it can be the *human's* reward: that is, it is possible to reward the state where the human drinks coffee, without also rewarding the state where the robot drinks coffee. This corresponds to the first principle: that machines should optimize *our* objectives. The second principle, that machines should initially be uncertain about our objectives, is

incorporated by assuming that *only H knows the reward*, requiring R to maintain a belief over the reward. Finally, for the third principle, R needs to get information about the reward from H's behavior, and so R assumes that H will choose actions that best optimize the reward (taking into account the fact that R doesn't know the reward).

This defines a two-player game, originally called a CIRL game but now called an *assistance game*. We can compute optimal joint strategies for H and R. Since this is an *interactive* process, H can do better than just acting optimally as if R did not exist (the assumption typically made in IRL): H can *teach* R what the reward is. In addition, R does not simply passively listen and then act, but interleaves learning and acting, and so must manage the explore-exploit tradeoff.

See also Learning to Interactively Learn and Assist (AN #64), which is inspired by this paper and does a similar thing with deep RL.

Read more: BAIR blog post: Cooperatively Learning Human Values

The Off-Switch Game *(Dylan Hadfield-Menell et al)*: This paper studies theoretically the impact of uncertainty over the reward on R's incentives around potential off switches. It proposes the simplest model that the authors expect to lead to generalizable results. R and H are in an assistance game, in which R goes first. R may either take an action a, getting utility u, or shut itself down, getting utility 0. In either case, the game ends immediately. Alternatively, R can choose to wait, in which case H can either shut down R, getting utility 0, or allow R to go ahead with action a, getting utility u.

If H is perfectly rational, then waiting is always an optimal action for R, since H will ensure that the team gets max(u, 0) utility. There can be other optimal actions: if R is sure that u >= 0, then taking action a is also optimal, and similarly if R is sure that u <= 0, then shutting down is also optimal. However, if H is not rational, and sometimes fails to take the utility-maximizing action (in a way R can't predict), then things get murkier. If R is sure about the value of u, then it is never optimal to wait, better to just take the action a (if u >= 0) or shut down (if u < 0) rather than let H screw it up. If R is pretty confident that u is positive, it may still decide to take action a, rather than risk that H makes the wrong decision. However, if R is very uncertain about the sign of u, then waiting becomes optimal again. In general, more uncertainty over the reward leads to more deferential behavior (allowing H to shut it off), but at a cost: R is much less able to help H when it is very uncertain about the reward.

Rohin's opinion: While I agree with the broad thrust of this paper, I do have one nitpick: the game ends immediately after H chooses whether or not to shut off R. In reality, if R isn't shut off, the assistance game will continue, which changes the incentives. If R can be relatively confident in the utility of *some* action (e.g. doing nothing), then it may be a better plan for it to disable the shutdown button, and then take that action and observe H in the mean time to learn the reward. Then, after it has learned more about the reward and figured out why H wanted to shut it down, it can act well and get utility (rather than being stuck with the zero utility from being shut down). While this doesn't seem great, it's not *obviously* bad: R ends up doing nothing until it can figure out how to actually be useful, hardly a catastrophic outcome. Really bad outcomes only come if R ends up becoming confident in the wrong reward due to some kind of misspecification, as suggested in Incorrigibility in the CIRL Framework, summarized next.

Incorrigibility in the CIRL Framework *(Ryan Carey)*: This paper demonstrates that when the agent has an *incorrect* belief about the human's reward function, then you no

longer get the benefit that the agent will obey shutdown instructions. It argues that since the purpose of a shutdown button is to function as a safety measure of last resort (when all other measures have failed), it should not rely on an assumption that the agent's belief about the reward is correct.

Rohin's opinion: I certainly agree that if the agent is wrong in its beliefs about the reward, then it is quite likely that it would not obey shutdown commands. For example, in the off switch game, if the agent is incorrectly certain that u is positive, then it will take action a, even though the human would want to shut it down. See also these (AN #32) posts (AN #32) on model misspecification and IRL. For a discussion of how serious the overall critique is, both from HC's perspective and mine, see the opinion on the next post.

Problem of fully updated deference *(Eliezer Yudkowsky)*: This article points out that even if you have an agent with uncertainty over the reward function, it will acquire information and reduce its uncertainty over the reward, until eventually it can't reduce uncertainty any more, and then it would simply optimize the expectation of the resulting distribution, which is equivalent to optimizing a known objective, and has the same issues (such as disabling shutdown buttons).

Rohin's opinion: As with the previous paper, this argument is only really a problem when the agent's belief about the reward function is *wrong*: if it is correct, then at the point where there is no more information to gain, the agent should already know that humans don't like to be killed, do like to be happy, etc. and optimizing the expectation of the reward distribution should lead to good outcomes. Both this and the previous critique are worrisome when you can't even put a reasonable *prior* over the reward function, which is quite a strong claim.

HC's response is that the agent should never assign zero probability to any hypothesis. It suggests that you could have an expandable hierarchical prior, where initially there are relatively simple hypotheses, but as hypotheses become worse at explaining the data, you "expand" the set of hypotheses, ultimately bottoming out at (perhaps) the universal prior. I think that such an approach could work in principle, and there are two challenges in practice. First, it may not be computationally feasible to do this. Second, it's not clear how such an approach can deal with the fact that human preferences *change* over time. (HC does want more research into both of these.)

Fully updated deference could also be a problem if the *observation model* used by the agent is incorrect, rather than the prior. I'm not sure if this is part of the argument.

Inverse Reward Design *(Dylan Hadfield-Menell et al)*: Usually, in RL, the reward function is treated as the *definition* of optimal behavior, but this conflicts with the third principle, which says that human behavior is the ultimate source of information about human preferences. Nonetheless, reward functions clearly have some information about our preferences: how do we make it compatible with the third principle? We need to connect the reward function to human behavior somehow.

This paper proposes a simple answer: since reward designers usually make reward functions through a process of trial-and-error where they test their reward functions and see what they incentivize, the reward function *tells us about optimal behavior in the **training** environment(s)*. The authors formalize this using a Boltzmann rationality model, where the reward designer is more likely to pick a *proxy reward* when it gives higher *true reward* in the *training environment* (but it doesn't matter if the proxy

reward becomes decoupled from the true reward in some test environment). With this assumption connecting the human behavior (i.e. the proxy reward function) to the human preferences (i.e. the true reward function), they can then perform Bayesian inference to get a posterior distribution over the *true* reward function.

They demonstrate that by using risk-averse planning with respect to this posterior distribution, the agent can avoid negative side effects that it has never seen before and has no information about. For example, if the agent was trained to collect gold in an environment with dirt and grass, and then it is tested in an environment with lava, the agent will know that even though the specified reward was indifferent about lava, this doesn't mean much, since *any* weight on lava would have led to the same behavior in the training environment. Due to risk aversion, it conservatively assumes that the lava is bad, and so successfully avoids it.

See also Active Inverse Reward Design (AN #24), which builds on this work.

Rohin's opinion: I really like this paper as an example of how to apply the third principle. This was the paper that caused me to start thinking about how we should be thinking about the assumed vs. actual information content in things (here, the key insight is that RL typically assumes that the reward function conveys much more information than it actually does). That probably influenced the development of Preferences Implicit in the State of the World (AN #45), which is also an example of the third principle and this information-based viewpoint, as it argues that the state of the world is caused by human behavior and so contains information about human preferences.

It's worth noting that in this paper the lava avoidance is both due to the belief over the true reward, *and the risk aversion*. The agent would also avoid pots of gold in the test environment if it never saw it in the training environment. IRD only gives you the correct uncertainty over the true reward; it doesn't tell you how to use that uncertainty. You would still need safe exploration, or some other source of information, if you want to reduce the uncertainty.

Prediction markets for internet points?

Using real money in prediction markets is all-but-illegal, and dealing with payments is a pain. But using fake money in prediction markets seems tricky, because by default players have no skin in the game.

Here's a simple proposal that I think might work reasonably well without being too hard to try:

- Create a service that tracks Internet Points. Anyone can quickly sign up with a Facebook account and is initially given 1000 points. Points are non-transferrable and are permanently associated with your Facebook account.
- Run prediction markets in which people wager points rather than dollars. You can initially copy questions and verdicts from existing markets and bookies. (Ideally you can just fork an existing market implementation.
- Keep the point tracking simple and transparent. Allow anyone to observe anyone's point total, and the history of bets that led to those totals. Avoid issuing new points. Try to create a sense of stability. No leaderboards, just observing your friends' (or strangers') point totals. (Maybe allow people to deactivate their account to hide their point total page, but if they reactivate they still have their old point total.)
- Limit maximum exposure to 50% of net worth. (But people can "exit" a market by taking a bet on the other side—betting on a 5 year question doesn't necessarily tie up your points indefinitely.) Encourage people to optimize their log returns rather than expected returns.
- Make it easy to link to pages for particular markets so people can see the current spread on a given question; make it very easy to quickly create a new account and bet on that market (2 clicks—login with facebook, and allow service to access your profile info).

It may be that a close-enough service already exists, though I'm a bit skeptical.

If you are at all interested in doing this, you have my blessing, this is a pretty generic idea (similar to sites that have existed in the past, e.g. Foresight Exchange). I'd also be happy to pay a bounty for an implementation I think is reasonable, I'd pay at least $1k for something that seems OK and up to $10k for something I think is actually really good.

Why I think this would be good

The barriers to entry in these markets would be tiny compared to existing prediction markets—within 30 seconds of seeing someone mention a market on Facebook, I could have created an account and placed a bet. I would likely bet in this kind of market if a few of my friends did.

Some (though far from all) people would take these markets seriously. For events in the next year or so, I think it's likely that these markets would quickly converge to better predictions than existing political prediction markets—if prices stayed crazy I think there are plenty of people who would be excited to step in to fix them even if they didn't take Internet Points very seriously (after all, they *might* be worth some

reputation in the future, the cost is super low, and betting on things is fine). Note that people who step in early to correct mispricings would be able to exit the market after sanity was restored, so wouldn't even need to tie up their points very long, and that they would have an incentivize to advertise the "easy win" to others after they'd bet in order to free up their capital.

If Internet Points are simple enough for people to really understand, and people expect them to stick around, I think they could potentially capture a big chunk of the altruistic upside from medium-stakes prediction markets.

If the overall workflow is very smooth, I think there is a reasonable chance that the site could spread quite rapidly (spreading just as well to new users as to people who have already made an account).

Note that playing this game can still be a fine experience even if you've lost most of your points—going from 250 to 1000 points is no harder than going from 1000 to 4000, it's just that the numbers are smaller.

Fake accounts and private bets

Ideally you'd have exactly 1 account per person. But creating Facebook accounts is quite cheap. Since every account is granted 1000 points, that's a problem. Some bad things that could happen, and what you might do about it:

- Someone could lose some bets, and then make a new account. But then their points won't be associated with their identity (and most people wouldn't start a new FB account to get more internet points), so as long as you mostly care about "your" points then this isn' ta big problem.
- Someone could make fake accounts and try to take their points. To avoid this, you probably want to make it impossible to transfer points except in public markets—the worst you can do is have some sock puppet accounts make bad bets and then take the other side, but if we structure markets appropriately this might be hard to do discretely. We can also make it against the terms of service and have simple mechanisms to try to detect it (e.g. by disallowing new FB accounts until we have a better system in place).
- Someone could make fake accounts and use them to influence markets, if they care about the market odds (e.g. if they want their preferred political candidate to look like they have a better chance in the general election). This can be partly addressed by simple measures above (like prohibiting new FB accounts).

It would be nice to fix these problems in a more robust way, so that market odds are more trustworthy, we can allow private bets, *etc.* That seems like a big project, over the long run I think the principled solution is basically a credit network: I indicate that I'm willing to trust my friends for IOUs, and then all bets need to be implemented as a string of IOU exchanges. This means that the social network can eventually fragment into different pieces (with the fake accounts basically living in their own part of the social network, which has very few trust connections with real accounts), and the market odds will potentially be different for different people. That makes everything way more subtle to think about, and the implementation seems *way* more complex, but it seems to me like it should work over the long run and is where you'd really like to be.

(You could also integrate that system with identity verification services who receive a high degree of trust, and then a service itself could effectively be deactivated—have

all of its trust exhausted—if it certified a bunch of users who lost money.)

Subsidies

A basic problem with this mechanism is that there is no subsidy (except for noise traders). It's not as bad as existing prediction markets, where the house takes a huge cut, but not as good as most "real" applications of prediction markets where someone interested in the info is willing to pay for it.

If I'm betting with someone reasonable, at least one of us is losing expected log-points (since the log bakes in risk aversion). So the market can't work with participants who trust each other's rationality, unless people just love gambling.

The most natural way to fix this is to subsidize markets. But doing this makes the governance and interpretation of Internet Points harder—does someone with 4000 points have a great history of losing bets, or did they just collect a subsidy? who are we effectively subsidizing, and how do we make that call?—so I'd prefer avoid it. If necessary, probably the nicest way to do it would be for the house to provide a modest amount of liquidity at the prices implied by existing markets. Existing prediction markets are sufficiently irrational that I think this would provide a significant incentive to participate.

If prices on Internet Point markets are used seriously as forecasts, it also introduces an additional set of incentives—some people want to manipulate forecasts (e.g. to make it look like their preferred candidate would be more likely to win the general election). So the average truth-seeking trader expects to get money by betting against them, and this can make the whole system work.

In some sense this is actually the only case where you really needed prediction markets anyway—if everyone is being reasonable then we can just talk it out and share our predictions, and that naive mechanism only really breaks down when some participants believe other participants are being predictably unreasonable (for whatever reason). For better or worse, this is a pretty typical situation.

Thoughts on "Human-Compatible"

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

> The purpose of this book is to explain why [superintelligence] might be the last event in human history and how to make sure that it is not... The book is intended for a general audience but will, I hope, be of value in convincing specialists in artificial intelligence to rethink their fundamental assumptions.

Yesterday, I eagerly opened my copy of Stuart Russell's *Human Compatible* (mirroring his *Center for Human-Compatible AI*, where I've worked the past two summers). I've been curious about Russell's research agenda, and *also* how Russell argued the case so convincingly as to garner the following acclamations from two Turing Award winners:

> Human Compatible made me a convert to Russell's concerns with our ability to control our upcoming creation—super-intelligent machines. Unlike outside alarmists and futurists, Russell is a leading authority on AI. His new book will educate the public about AI more than any book I can think of, and is a delightful and uplifting read.—Judea Pearl

> This beautifully written book addresses a fundamental challenge for humanity: increasingly intelligent machines that do what we ask but not what we really intend. Essential reading if you care about our future. —Yoshua Bengio

Bengio even recently lent a reasoned voice to a debate on instrumental convergence!

Bringing the AI community up-to-speed

I think the book will greatly help AI professionals understand key arguments, avoid classic missteps, and appreciate the serious challenge humanity faces. Russell straightforwardly debunks common objections, writing with both candor and charm.

I must admit, it's great to see such a prominent debunking; I still remember, early in my concern about alignment, hearing one professional respond to the entire idea of being concerned about AGI with a lazy *ad hominem* dismissal. Like, hello? This is our future we're talking about!

But Russell realizes that most people don't intentionally argue in bad faith; he structures his arguments with the understanding and charity required to ease the difficulty of changing one's mind. (Although I wish he'd be a little less sassy with LeCun, understandable as his frustration may be)

More important than having fish, however, is knowing how to fish; Russell helps train the right mental motions in his readers:

> With a bit of practice, you can learn to identify ways in which the achievement of more or less any fixed objective can result in arbitrarily bad outcomes. [Russell

goes on to describe specific examples and strategies] (p139)

He somehow explains the difference between the Platonic assumptions of RL and the reality of a human-level reasoner, *while also* introducing wireheading. He covers the utility-reward gap, explaining that our understanding of real-world agency is so crude that we can't even coherently talk about the "purpose" of eg AlphaGo. He explains instrumental subgoals. These bits are so, so good.

Now for the main course, for those already familiar with the basic arguments:

The agenda

Please realize that I'm replying to my understanding of Russell's agenda as communicated in a nontechnical book for the general public; I also don't have a mental model of Russell personally. Still, I'm working with what I've got.

Here's my summary: reward uncertainty through some extension of a CIRL-like setup, accounting for human irrationality through our scientific knowledge, doing aggregate preference utilitarianism for all of the humans on the planet, discounting people by how well their beliefs map to reality, perhaps downweighting motivations such as envy (to mitigate the problem of everyone wanting positional goods). One challenge is towards what preference-shaping situations the robot should guide us (maybe we need meta-preference learning?). Russell also has a vision of many agents, each working to reasonably pursue the wishes of their owners (while being considerate of others).

I'm going to simplify the situation and just express my concerns about the case of one irrational human, one robot.

There's fully updated deference:

> One possible scheme in AI alignment is to give the AI a state of moral uncertainty implying that we know more than the AI does about its own utility function, as the AI's meta-utility function defines its ideal target. Then we could tell the AI, "You should let us shut you down because we know something about your ideal target that you don't, and we estimate that we can optimize your ideal target better without you."
>
> The obstacle to this scheme is that belief states of this type also tend to imply that an even better option for the AI would be to learn its ideal target by observing us. Then, having 'fully updated', the AI would have no further reason to 'defer' to us, and could proceed to directly optimize its ideal target.

which Russell partially addresses by advocating ensuring realizability, and avoiding feature misspecification by (somehow) allowing for dynamic addition of previously unknown features (see also *Incorrigibility in the CIRL Framework*). But supposing we don't have this kind of model misspecification, I don't see how the "AI simply fully computes the human's policy, updates, and then no longer lets us correct it" issue is addressed. If you're really confident that computing the human policy lets you just extract the true preferences under the realizability assumptions, maybe this is fine? I suspect Russell has more to say here that didn't make it onto the printed page.

There's also the issue of getting a good enough human mistake model, *and* figuring out people's beliefs, all while attempting to learn their preferences (see the value learning sequence).

Now, it would be pretty silly to reply to an outlined research agenda with "but specific problems X, Y, and Z!", because the whole point of further research is to solve problems. However, my concerns are more structural. Certain AI designs lend themselves to more robustness against things going wrong (in specification, training, or simply having fewer assumptions). It seems to me that the uncertainty-based approach is quite demanding on getting component after component "right enough".

Let me give you an example of something which is intuitively "more robust" to me: approval-directed agency.

> Consider a human Hugh, and an agent Arthur who uses the following procedure to choose each action:
>
> ```
> Estimate the expected rating Hugh would give each action if he considered it at
> length. Take the action with the highest expected rating.
> ```

Here, the approval-policy does what a predictor says to do at each time step, which is different from maximizing a signal. Its *shape* feels different to me; the policy isn't *shaped* to maximize some reward signal (and pursue instrumental subgoals). Errors in prediction almost certainly don't produce a policy adversarial to human interests.

How does this compare with the uncertainty approach? Let's consider one thing it seems we need to get right:

Where in the world is the human?

How will the agent robustly locate the human whose preferences it's learning, and why do we need to worry about this?

Well, a novice might worry "what if the AI doesn't properly cleave reality at its joints, relying on a bad representation of the world?". But, having good predictive accuracy is instrumentally useful for maximizing the reward signal, so we can expect that its implicit representation of the world continually improves (i.e., it comes to find a nice efficient encoding). We don't have to worry about this - the AI is incentivized to get this right.

However, if the AI is meant to deduce and further the preferences of that single human, it has to find that human. But, before the AI is operational, how do we point to *our* concept of "this person" in a yet-unformed model whose encoding probably doesn't cleave reality along those same lines? Even if we fix the structure of the AI's model so we *can* point to that human, it might then have instrumental incentives to modify the model so it can make better predictions.

Why does it matter so much that we point exactly to the human? Well, then we're extrapolating the "preferences" of something that is not the person (or a person?) - the predicted human policy in this case seems highly sensitive to the details of the person or entity being pointed to. This seems like it could easily end in tragedy, and (strong belief, weakly held) doesn't seem like the kind of problem that has a clean solution. this sort of thing seems to happen quite often for proposals which hinge on things-in-ontologies.

Human action models, mistake models, etc. are also difficult in this way, and we have to get them right. I'm not necessarily worried about the difficulties themselves, but that the framework seems so sensitive to them.

Conclusion

This book is most definitely an important read for both the general public and AI specialists, presenting a thought-provoking agenda with worthwhile insights (even if I don't see how it all ends up fitting together). To me, this seems like a key tool for outreach.

Just think: in how many worlds does alignment research benefit from the advocacy of one of the most distinguished AI researchers ever?

Theater Tickets, Sleeping Pills, and the Idiosyncrasies of Delegated Risk Management

Risk management is difficult, but even when it’s easy, companies and policymakers often do something other than optimal risk mitigation. This isn’t puzzling, once we realize that the incentives in place give the decisionmakers the leeway, or even positive incentives, to behave sub-optimally. There are three types that seem most relevant, along with a few (anonymized) stories from when I was working in reinsurance of how they play out in practice.

Sleeping Pills

Occasionally, a small insurance company would purchase reinsurance for things that didn’t make business sense for their company. I might have seen a home insurance company in Ohio that had fifty million dollars in reserve, then would buy reinsurance for hurricanes that covered all losses greater than ten and less than twenty five million dollars. Yes, there are hurricanes that impact Ohio, such as Xenia in 1974 and Ike in 2008, but they weren’t large enough to even hit the minimum for this type of policy. Not only that, but the company had money in reserves to cover this incredibly unlikely loss, and buying reinsurance isn't cheap. Let’s say that between brokers, transaction costs, and everything else, it cost them a hundred thousand dollars to cover an expected loss of ten thousand dollars.

Noticing this, I asked why they wanted this policy. My boss told me it was a sleeping pill. He explained that the CEO of the insurance company would get really nervous and unhappy every time a hurricane was approaching the US, and decided this CEO didn’t want to worry anymore. That isn’t unreasonable - most people who buy travel insurance to cover their $1,000 vacation could just accept the risk, but they prefer not to worry.

In general, buying risk mitigation isn’t worth the cost in expected value terms, but they are worthwhile because they can buy off the worry. In this case, it’s less innocuous, since the CEO was using company money to buy what amounted to personal sleeping pills. It happens because the CEO won’t ever be blamed for hedging the risk, and it cost the company a hundred thousand dollars per year, which was not enough for shareholders to notice.

Theater Tickets

Once, we received a request from a client to train them in using the terrorism risk model they licensed. That’s not unusual, but they said that they wanted us to first come and install the software, then do a half day training. This seemed weird, since they had been licensing the software for several years, and we assumed they would have already installed it. They hadn’t.

The software licenses weren’t cheap - I don’t remember the client or the amount, but it was easily a six-figure sum every year. Why did they pay if they didn’t even use the model? It was what Bruce Schnier calls “Security Theater,” referring to acts that do

nothing to change the risk, but look good. In this case, they wanted to tell their shareholders in their annual report that they had a model for assessing their terrorism risk, and six-figures was cheap to be able to put on the show of risk-management and mitigation for their shareholders. In this case, they didn't even need to put on a show - just waving the tickets they bought was enough. They paid money not to mitigate risk, but simply to make it look like they did.

Staying Out of the Pool

There's another phenomenon where managers don't want to take risks that are good for the company but risky for their division - and their careers. The typical version is explained in a 1966 article in Harvard Business Review, "Utility theory, insights into risk taking." They asked managers if they'd take a 50-50 chance on a project that would either make $300,000, or lose $60,000, and the managers mostly said no.

The reason this is a bad decision is that even if the company prefers to avoid large risks, the company should want middle-managers to be taking smaller risks. They want small risks because there are lots of them, and in aggregate - by pooling the risks - the company is far better off having managers take them. For a single bet, the expected return is $120,000, but there's a 50% chance of not only not coming out even, but actually losing money. If 6 managers took similar (uncorrelated) bets, the expected value is six times as large - it scales linearly - but the probability of losing money in that case is $(\frac{1}{2})^6$, or under 2%.

Reinsurance is actually a place that understands this far better than most. Since they explicitly model the risks of the insurance contracts, and the entire reason insurance works is risk pooling, they are far better equipped to handle the problem. Still, underwriters get paid their bonus based on their performance, one that likely won't materialize if they are net negative for the year. So even here, you see a hesitation for the individuals to get into the (risk) pool.

Conclusion

None of these are new insights or issues. All of them are simple combinations of a principal-agent issue and risk aversion. Still, in addition to reinforcing the ideas, they are worth thinking about when we have one person or group of people mitigate risks for others. The obvious places are in the corporate world, and in government, and if you look around, you'll see that these dynamics are all common.

Climate technology primer (1/3): basics

This is my first post here, cross-posted from

https://longitudinal.blog/co2-series-part-1-review-of-basics/

Posting was suggested by Vaniver, and the actual work of cross-posting was done by Benito.

Comments appreciated. (Will only cross post the first of the 3 climate posts here.)

This is the first of a series of three blog posts intended as a primer on how technology can help to address climate change.

Note: you can annotate this page in Hypothes.is here: https://via.hypothes.is/https://longitudinal.blog/co2-series-part-1-review-of-basics/

Excellent documents and courses are available online on what we are doing to the climate — such as the IPCC reports or Prof. Michael Mann's online course or this National Academies report. Even more so, this type of open source analysis is what the field needs to look like. But I had to dig a bit to see even the beginnings of a quantitative picture of how technologies can contribute to *solving* the problems.

There are integrative assessments of solutions, such as Project Drawdown, and a huge proliferation of innovative technologies being proposed and developed in all sectors. There are also some superb review papers that treat the need for new technologies. Bret Victor has a beautiful overview emphasizing the role of general design, modeling and implementation capacity, and there is now a climate change for computer scientists course at MIT. But, as a scientifically literate (PhD in Biophysics) *outsider* to the climate and energy fields, I was looking to first understand the very *basics*:

- How can I estimate the world's remaining carbon budget for various levels of warming? What are the uncertainties in this?
- What would happen if Greenland melted?
- What are the fundamental physical limits on, say, the potential for sucking carbon dioxide out of the atmosphere? How low could the cost theoretically go?
- How many trees would you need to plant, and are you going to run out of water, land or fertilizer for them?
- If we did manage to remove a given amount of CO2, how would the climate respond? How would the climate respond if we magically ceased all emissions today?
- What would be the real impact if we could electrify all cars?
- Are batteries going to be able to store all the variable renewable energy we'll need? Is there enough Uranium obtainable to power the world with nuclear? Order of magnitude, how much would it cost to just replace our entire energy supply with nuclear?
- Notwithstanding the complexity of their governance and potential moral hazard implications, do we have a solid science and engineering basis to believe that any geoengineering methods could theoretically work to *controllably* lower temperature, and if so, would that be able to slow down sea ice melt, too? Would it damage the ozone layer? How fast could you turn it off? Could it be done locally?
- Etc.

For questions like those and many others, I want to calculate "on the back on an envelope" an estimate of whether any given idea is definitely crazy, or might just be possible.

I suspect that a similar desire led the information theory and statistics luminary David MacKay to write his seminal calculation-rich book Sustainable Energy without the Hot Air.

While these blog posts, written over a few vacations and weekends by a more modest intellect, don't begin to approach MacKay's standard of excellence, I do hope that they can serve as a breadcrumb trail through the literature that may help others to apply this spirit to a broader set of climate-relevant engineering problems.

In this first of three posts, I attempt an outsider's summary of the basic physics/chemistry/biology of the climate system, focused on *back of the envelope calculations* where possible. At the end, I comment a bit about technological approaches for emissions reductions. Future posts will include a review of the science behind negative emissions technologies, as well as the science (with plenty of caveats, don't worry) behind more controversial potential solar radiation management approaches. This first post should be very basic for anyone "in the know" about energy, but I wanted to cover the basics before jumping into carbon sequestration technologies.

(These posts as a whole are admittedly weighted more towards what I perceived to be less-mainstream topics, like carbon capture and geo-engineering. In this first one, while I do briefly cover a lot of cool mitigation techniques — like next-generation renewables storage, a bit about nuclear, and even a bit about the economics of carbon taxes — I don't try to cover the full spectrum of important technology areas in mitigation and adaptation. But don't let that leave you with the impression that the *principal* applications of tech to climate are in CO2 removal and geo-engineering. Areas I don't cover much here may be *even more important and disruptive* — like, to name a few, decarbonizing steel and cement production, or novel control systems for solar and wind power, or the use of buildings as smart grid-adaptive power loads, or hydrogen fuel, or improved weather prediction… It just happens that the scope of my *particular* review here is: some traditional mitigation, a lot of CDR and geo-engineering, and ~nothing in adaptation aspects like public health or disaster relief. Within applications of machine learning to climate, I recommend this paper for a broader picture that spans a wider range of solution domains — such breadth serves as good food for thought for technologists interested in finding the truly most impactful areas in which they might be able to contribute, but I don't attempt it here.)

Disclaimers:

1. I am not a climate scientist or an energy professional. I am just a scientifically literate lay-person on the internet, reading up in my free time. If you are looking for some climate scientists online, I recommend those that Michael Nielsen follows, and this list from Prof. Katharine Hayhoe. Here are some who apparently advise Greta. You should definitely read the IPCC reports, as well, and take courses by real climate scientists, like this one. The National Academies and various other agencies also often have reports that can claim a much higher degree of expertise and vetting than this one.
2. Any views expressed here are mine only and do not reflect any current or former employer. Any errors are also my own. This has not been peer-reviewed in any meaningful sense. If you're an expert on one of these areas I'll certainly appreciate and try to incorporate your suggestions or corrections.
3. There is not really anything new here — I try to closely follow and review the published literature and the scientific consensus, so at most my breadcrumb trail may serve to make you more aware of what scientists have already told the world in the form of publications.
4. I'm focused here on trying to understand the narrowly technical aspects, not on the political aspects, despite those being crucial. This is meant to be a review of the technical literature, not a political statement. I worried that writing a blog purely on the topic of technological intervention in the climate, without attempting or claiming to do justice to the social issues raised, would implicitly suggest that I advocate a narrowly technocratic or unilateral approach, which is not my intention. By focusing on technology, I don't mean to detract from the importance of the social and policy aspects. I do mention the importance of carbon taxes several times, as possibly necessary to drive the development and adoption of technology. I don't mean to imply through my emphasis that all solutions are technologically advanced — for example,

crucial work is happening on conservation of land and biodiversity. That said, I do view advanced technology as a key lever to allow solutions to scale worldwide, at the hundreds-of-gigatonnes-of-CO2 level of impact, in a cost-effective, environmentally and societally benign way. Indeed, the rights kinds of improvements to our energy system are likely one of the best ways to spur economic growth.
5. Talking about emerging and future technologies doesn't mean we shouldn't be deploying existing decarbonization technologies now. There is a finite cumulative carbon budget to avoid unacceptable levels of warming. A perfect technology that arrives in 2050 doesn't solve the primary problem.
6. For some of the specific technologies discussed, I will give further caveats and arguments in favor of caution in considering deployment.

Acknowledgements: I got a bunch of good suggestions from friends including Brad Zamft, Sarah Sclarsic, Will Regan, David Pfau, David Brown, David Rolnick, Tom Hunt, Tony Pan, Marcus Sarofim, Michael Nielsen, Sam Rodriques, James Ough, Evan Miyazono, Nick Barry, Kevin Esvelt, Eric Drexler and George Church.

Outline:

Basic numbers on climate

Let's start with crude "back of the envelope" approximations — of a type first done by Svante Arrhenius in 1896, who got the modern numbers basically right — and see how far we get towards understanding global warming (the phenomenon of CO2-driven warming itself was arguably discovered around 1856 by Eunice Foote). We'll quickly hit a complexity barrier to this approach, but the lead-up was interesting to me.

Radiation balance

The atmosphere is complicated, many layered, and heterogeneous. But as a quick tool for thought, we can start with simple **single-layer atmosphere models**, aiming to understand the very basics of the effect of the atmosphere on the Earth's balance of incoming sunlight versus outgoing thermal infrared radiation. These basics are covered at the following excellent sites

- acs.org/content/acs/en/climatescience/atmosphericwarming/singlelayermodel.html
- web.gps.caltech.edu/classes/ese148a/ especially lecture 2
- mathaware.org/mam/09/essays/Radiative_balance.pdf
- clivebest.com/blog/?p=2241
- assets.press.princeton.edu/chapters/s9636.pdf
- https://geosci.uchicago.edu/~rtp1/papers/PhysTodayRT2011.pdf
- https://en.m.wikipedia.org/wiki/Idealized_greenhouse_model
- https://twitter.com/michael_nielsen/status/1096990267259809793
- https://www.e-education.psu.edu/meteo469/node/198

which I now summarize.

An ideal black body radiator (e.g., a very idealized Planet Earth), with no greenhouse effect, would radiate an energy flux:

$$\text{no-greenhouse outgoing flux} = \sigma * T^4$$

where

$$\sigma = 5.67 * 10^{-8} \text{Watts/meter}^2/\text{Kelvin}^4$$

is the Stefan-Boltzmann constant, and T is the temperature in Kelvin.

If we multiply this flux by the surface area of the Earth, we have the outgoing emitted energy:

$$\text{no-greenhouse emitted energy} = 4 * \pi * (\text{radius of Earth})^2 * \sigma * T^4$$

If the planet was in radiative balance — so that the incoming solar energy is equal to the outgoing thermal energy and the planet would be neither warming nor cooling — we'd want to set this equal to the incoming solar energy

$$\text{incoming solar energy} = S * (1 - A) * \pi * (\text{radius of Earth})^2$$

This last equation introduced the *albedo* A, which is the fraction of incoming solar energy reflected directly back without being absorbed, and the incoming solar flux S from the sun.

The incoming solar radiation is roughly S = 1.361 kilowatts per square meter, and the albedo A is around 0.3. S is easy to calculate: the sun has a radius of 700 million meters and a temperature of 5,778K, so the energy radiated by the sun is, by the same formula we used for the Earth,

$$\sigma * 4 * \pi * (700 \text{ million meters})^2 * (5778 \text{ Kelvin})^4$$

which is then by the time it reaches the Earth spread over a sphere of radius given by the Sun-Earth distance of 1.5e11 meters, so the solar flux per unit area is

S =

$$\sigma * 4 * \pi * (700 \text{ million meters})^2 * (5778 \text{ Kelvin})^4 / (4 * \pi * (1.5e11 \text{ meters})^2)$$

= 1.376kW/meter^2.

(That's a lot of solar flux. If the area of Texas was covered by 4% efficient solar collection, that would offset the entire world's average energy consumption.)

We can then solve our incoming-outgoing solar energy balance for the single unknown, the equilibrium temperature T, which comes out to **255 Kelvin** or so. Now this is quite cold, below zero degrees Fahrenheit, whereas the actual global mean surface temperature is much warmer that, on average around 60 degrees Fahrenheit (288K or so) — **the reason for this difference is the greenhouse effect.**

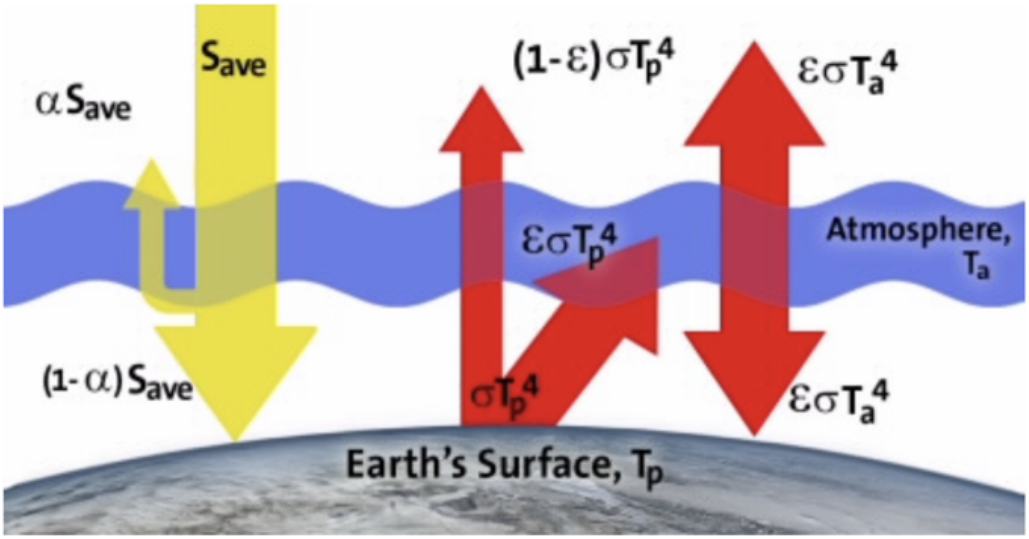
The greenhouse-gas-warmed situation that we find ourselves in is also simple to model at a crude level, but complex to model at a more accurate level.

At the crudest level (first few lectures of an undergraduate climate science course), if a certain fraction ε of outgoing infrared radiation is absorbed by the atmosphere on the way out (e.g., by its energy being converted into intra-molecular vibrations and rotations in the CO2 molecules in the atmosphere), and then re-emitted, then we can calculate the resulting ground temperature (a more complete explanation is in the lead-up to equation 1.15 here, or here), arriving at a ground temperature

$$T = [(2-f)*(1-A)*S/((2-\epsilon)*4*\sigma)]^{1/4}$$

where f is the fraction of incoming solar radiation absorbed by the atmosphere *before* making it to the surface, and ε is the so-called *atmospheric emissivity* — ε this is zero for no greenhouse gasses and 1 for complete absorption of infrared radiation from the surface by the atmosphere. I've seen f = 0.08 (the atmosphere absorbs very little, less than 10%, of the incoming solar radiation), and ε=0.89 although various other combinations do the same job. **If you plug this all in, you get a ground temperature of 286 Kelvin or so, which is pretty close to our actual global mean surface temperature of around 288 Kelvin.** (Note that even if ε=1, we don't have infinite warming, as the radiation is *eventually* re-radiated out to space.)

Here is the picture that goes with this derivation:



Credit: American Chemical Society

https://www.acs.org/content/acs/en/climatescience.html

Again, this is not a good model, because it neglects almost all the complexities of the atmosphere and climate system, but it can be used to illustrate *certain* basic things. For example, if you want to keep the Earth's temperature low, you can theoretically increase the albedo, A, to reflect a higher percentage of the incoming solar radiation back to space; this is called solar geoengineering or solar radiation management, and this formula allows to calculate crudely the kind of impact that can have.

But the formula is **hopelessly oversimplified** for most purposes, e.g., we've neglected the fact that much of the heat leaving the Earth's surface is actually due to evaporation, rather than direct radiation.

A significantly better framework, yet still fully analytical rather than simulation-based, is provided by this paper (and slides and these three videos), but the physics is more involved.

Climate sensitivity

A key further problem for us as back-of-the-envelope-ers, though, is to know what a given amount of CO2 in the atmosphere translates to in terms of *effective* ε and A (and f) in the above formula. Or in plain English, to relate the amount of atmospheric CO2 to the amount of warming. This is where things get trickier. Before trying to pin down a number for this "climate sensitivity", let's first review why doing so is a very complex and multi-factorial problem.

Water vapor

Naively, if we think of the only effect of CO2 as being the absorption of some of the outgoing infrared thermal radiation from the planet surface, then more CO2 means higher ε, with no change in albedo A. However, in reality, more warming also leads to more water vapor in the atmosphere, which — since **water vapor is itself a greenhouse gas** — can lead to more greenhouse effect and hence more warming, causing a *positive feedback*. On the other hand, if the extra water vapor forms clouds, that can increase the albedo, decreasing warming by blocking some of the incoming solar radiation from hitting the surface, and leading to a *negative feedback*. So it doesn't seem obvious even what functional form the impact of CO2 concentration on ε and A should take, particularly since water vapor can play multiple roles in the translation process.

If you just consider clouds per se, it looks like they on average cool the Earth, from an experiment described in lecture 10a of the Caltech intro course, but this depends heavily on their geographic distribution and properties, not to mention their dynamics like heating-induced evaporation, changes in large-scale weather systems, effects specific to the poles, and so on. **Aerosol-cloud interactions, and their effects on albedo and radiative forcing**, appear to be a notable area of limited scientific understanding, and the IPCC reports state this directly. Also, a particular type of cloud called Cirrus actually warms the Earth.

In general, the feedback effects of water are a key consideration and still imperfectly understood by the field, although the research is making great strides as we speak. For example, just looking at a fairly random sampling of recent papers, it looks like the nature of the water feedback may be such that the Stefan-Boltzmann law calculation above is no longer really appropriate even as an approximation, leading to a *linear* amount of outgoing radiation with temperature rather than *quartic* per the Stefan-Boltzmann law

pnas.org/content/115/41/10293

phys.org/news/2018-09-earth-space.html

Even weirder,

news.mit.edu/2014/global-warming-increased-solar-radiation-1110

pnas.org/content/111/47/16700

the roles of carbon dioxide and water may be such that the largest warming mechanism is increased *penetrance* of solar radiation rather than decreased *escape* of infrared per se. As the article explains:

> *"In computer modeling of Earth's climate under elevating CO2 concentrations, the greenhouse gas effect does indeed lead to global warming. Yet something puzzling happens: While one would expect the longwave radiation that escapes into space to decline with increasing CO2, the amount actually begins to rise. At the same time, the atmosphere absorbs more and more incoming solar radiation; it's this* ***enhanced shortwave absorption*** *that ultimately sustains global warming."*

Note that this paper is meant as an *explanation of known phenomena arising inside climate models*, **not** as a contradiction of those models.

So it is complicated. (I don't mean to over-emphasize these particular papers except to say that people are still discovering core aspects of how the Earth's energy balance works.)

Moreover, the feedback effects due to changes associated with water take a *long time* to manifest, due in part to the large specific heat of water and the large overall heat capacity of the enormous, deep oceans, and in part due to the long residence time of CO2 in the atmosphere. Thus, simply fitting curves for observed temperature versus CO2 concentration on a short timescale around the present time may fail to reflect **long-timescale processes** that will be kicking in over the coming decades, such as feedbacks.

Ocean biology effects

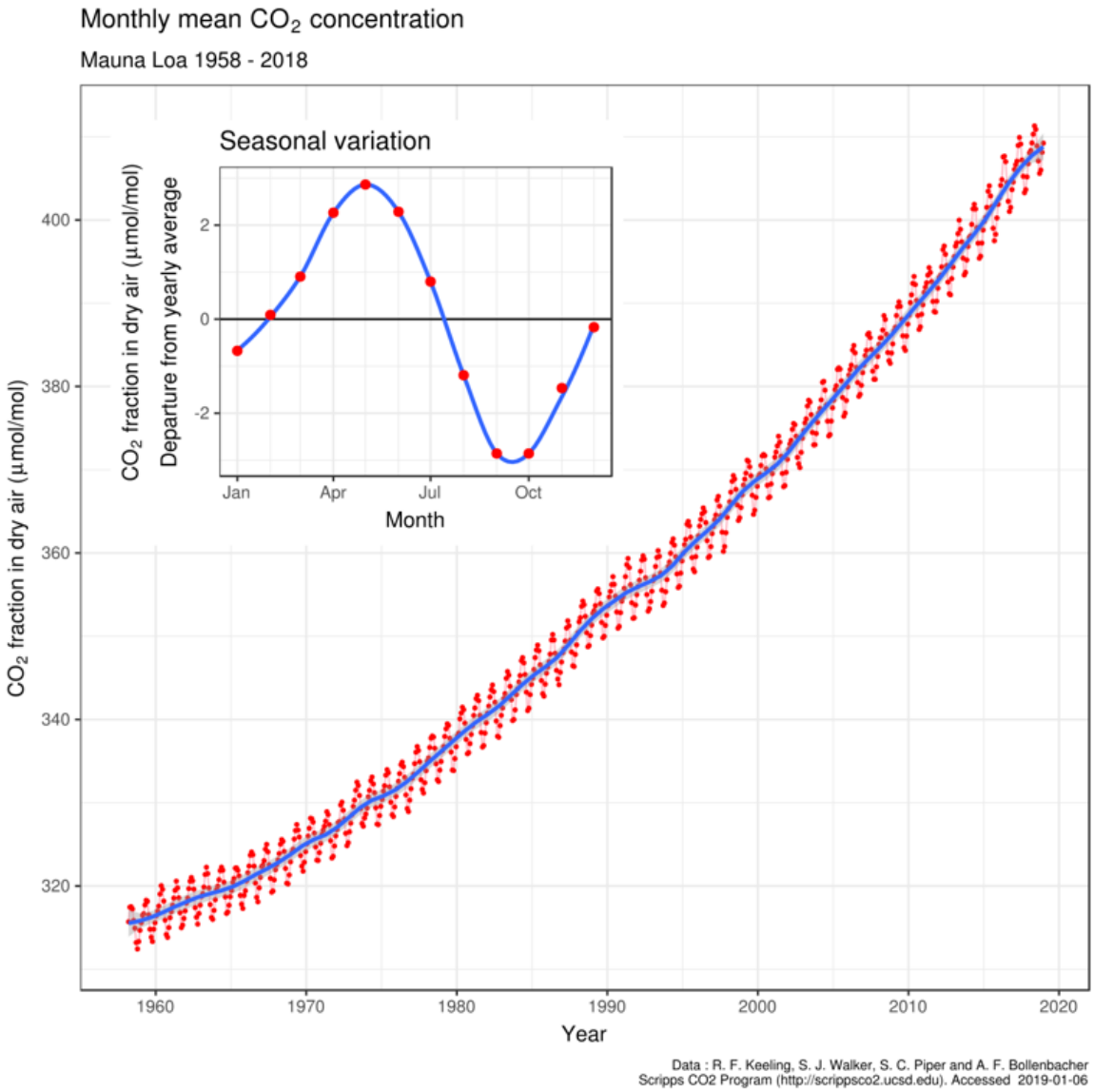
There are also large ocean biology and chemistry effects. Just to list a few, dissolved carbon impacts the acidity of the ocean, which in turn impacts its ability to uptake more carbon from the atmosphere. There are also major ocean *biological* effects from photosynthesis and the drivers of it. As the Caltech course explains:

> *"On longer time scales (decade to century), atmospheric CO2 is strongly influenced by mixing of deep ocean water to the surface. Although deep water is generally supersaturated with respect to even the modern atmosphere, mixing of these water masses to the surface tends to draw down atmospheric CO2 in the end as the high nutrients stimulate formation of organic carbon. Several coupled atmosphere-ocean models have shown, however, that global warming is accompanied by an increase in vertical stratification within the ocean and therefore reduced vertical mixing. On its own, this effect would tend to reduce the ocean CO2 uptake. However, changes in stratification may also drive changes in the natural carbon cycle. The* ***magnitude and even the sign of changes in the natural cycle are much more difficult to predict because of the complexity of ocean biological processes****…". Indeed, the Caltech slides invoke localized mixing effects in parts of the ocean, and their impact on ocean photosynthesis, as a possible explanation for major pre-industrial temperature changes on inter-glacial timescales.*

Land biology effects

There are also terrestrial biological effects. We'll use the great physicist Freeman Dyson's emphasis on these as a jumping off point to discuss them. (Dyson is known as a bit of a "heretic" on climate and many other topics. Here are 157 interviews with him. We won't dwell on his opinions about what's good for the world, just on the technical questions he

asked about the biosphere, because the history is cool.) In 1977, in a paper that will come up again for us in the context of carbon sequestration technologies, Dyson pointed out that the yearly photosynthetic turnover of CO2, with carbon going into the bodies of plants and then being released back into the atmosphere through respiration and decay, is >10x yearly industrial emissions. The biological turnover is **almost exactly in balance over a year, but not perfectly so at any location and instant of time**. The slight imbalances over time give the yearly oscillation in the Keeling curve:



Source: Wikipedia

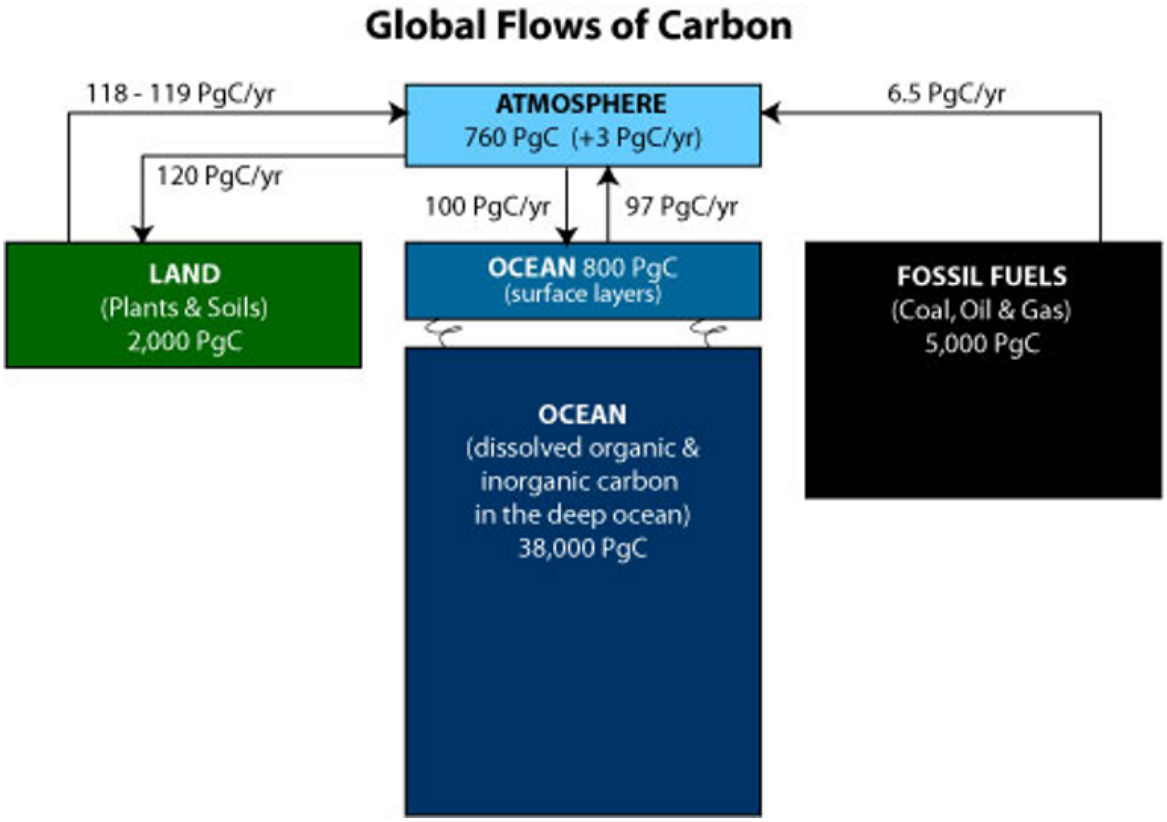
Recently, this paper measured the specifics, concluding "**150–175 petagrams of carbon per year**" of **photosynthesis**, whereas global emissions are roughly ~10 petagrams (gigatonnes) of carbon each year (~40 of CO2). (Here as everywhere I'm going to ignore the small difference between long and short tons and metric tons, i.e., tonnes). This is just for the land, while the oceans do about the same amount — see the diagram of the overall carbon cycle below. The influx into the ocean is in part just due to CO2 dissolving in the water, but a significant part of the long-term sequestration of carbon into the *deep* ocean is due to settling of biological material under gravity and other mechanisms.

Dyson stressed that any **net imbalances** that arise in this process, e.g., due to "CO2 fertilization" effects — which cause increased growth rate or mass of plant matter with higher available atmospheric CO2 concentration — could exert large effects. Note that biology may have once caused an ice age. Dyson goes on to discuss potential ways to tip the balance of this large swing towards *net* fixation, rather than full turnover, which he suggests to do via a plant growing program — we'll discuss this in relation to biology-based carbon sequestration concepts later on.

Measurements of CO2 fertilization of land plants have improved since Dyson was writing, e.g., this paper, which gives a value: "…0.64 ± 0.28 PgC yr^{-1} per hundred ppm of eCO2. Extrapolating worldwide, this… *projects the global terrestrial carbon sink to increase by 3.5 ± 1.9 PgC yr^{-1} for an increase in CO2 of 100 ppm*". That is pretty small compared to emissions.

It seems like the size of the CO2 fertilization effect is limited by nutrient availability, as is much normal forest growth. The historical effect of CO2 fertilization on growth in a specific tree type is measured here, finding large impact on water use efficiency, but little on actual growth of mass. Another recent paper found that CO2 uptake by land plants may be facing limitations related to soil moisture variability, and "suggests that the increasing trend in carbon uptake rate may not be sustained past the middle of the century and could result in accelerated atmospheric CO2 growth".

This diagram (a bit out of date on the exact numbers) summarizes the overall pattern of ongoing fluxes in the carbon cycle



Source: NASA

and here is a nice animation of this over time as emissions due to human activity have rapidly increased.

Dyson bemoans our lack of knowledge of long-timescale ecological effects of CO2 changes on plant life, e.g., writing:

> *"Greenhouse experiments show that many plants growing in an atmosphere enriched with carbon dioxide react by increasing their root-to-shoot ratio. This means that the plants put more of their growth into roots and less into stems and leaves. A change in this direction is to be expected, because the plants have to maintain a balance between the leaves collecting carbon from the air and the roots collecting mineral nutrients from the soil.* ***The enriched atmosphere tilts the balance so that the plants need less leaf-area and more root-area.*** *Now consider what happens to the roots and shoots when the growing season is over, when the leaves fall and the plants die. The new-grown biomass decays and is eaten by fungi or microbes. Some of it returns to the atmosphere and some of it is converted into topsoil. On the average, more of the above-ground growth will return to the atmosphere and more of the below-ground growth will become topsoil. So the plants with increased root-to-shoot ratio will cause an increased transfer of carbon from the atmosphere into topsoil. If the increase in atmospheric carbon dioxide due to fossil-fuel burning has caused an increase in the average root-to-shoot ratio of plants over large areas, then the possible effect on the top-soil reservoir will not be small.* ***At present we have no way to measure or even to guess the size of this effect.*** *The aggregate biomass of the topsoil of the planet is not a measurable quantity. But the fact that the topsoil is unmeasurable does not mean that it is unimportant."*

Here is a paper on how we can indeed measure the soil carbon content. Dyson further discusses important measurements here (already 20 years ago), including on **biosphere-atmosphere fluxes**, and I believe one of the projects he is mentioning is at least somewhat similar to this one called FLUXNET.

Permafrost

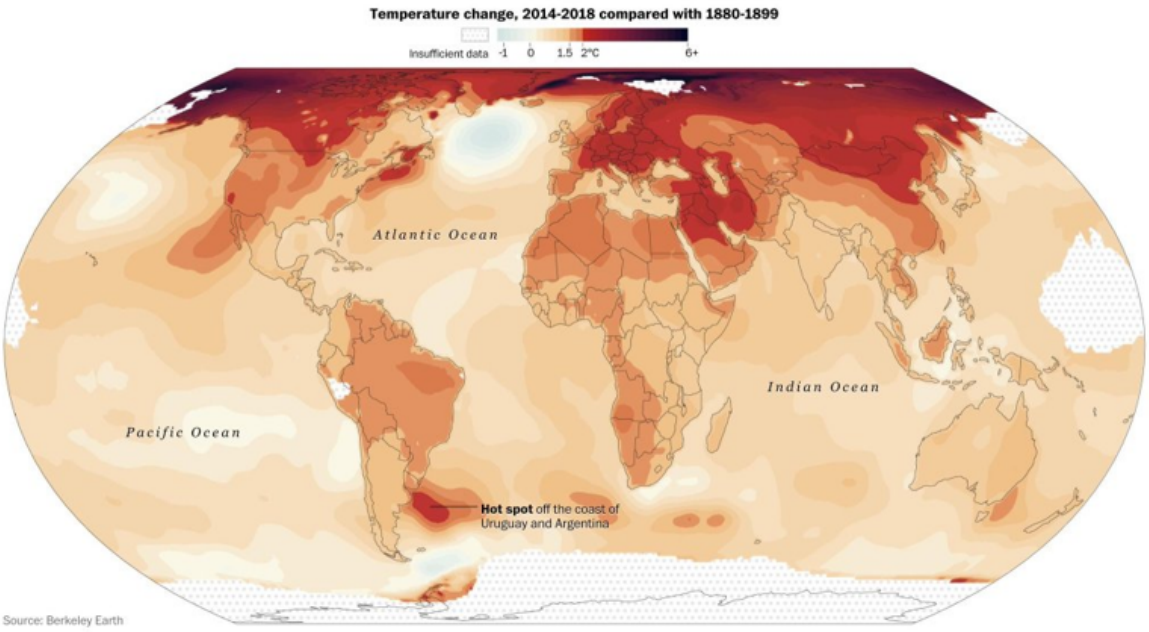
There are other large **potential** feedback effects, beyond those due to water vapor or plant life, e.g., over 1000 gigatonnes of carbon trapped in the arctic tundra, particularly in the form of the **greenhouse gas methane which exerts much stronger short term effects** than carbon, and which could be released due to thawing of the permafrost. (Eli Dourado covers the fascinating Pleistocene Park project which aims to influence this.) Wildfires are also relevant here. (Apparently recent Russian policy used to think of climate change as mostly a good thing until it became clear that the permafrost is at risk.)

Local warming

Dyson also emphasizes the importance of local measurements, for instance:

> *"In humid air, the effect of carbon dioxide on radiation transport is unimportant because the transport of thermal radiation is already blocked by the much larger greenhouse effect of water vapor. The effect of carbon dioxide is important where the air is dry, and air is usually dry only where it is cold. Hot desert air may feel dry but often contains a lot of water vapor. The warming effect of carbon dioxide is strongest where air is cold and dry, mainly in the arctic rather than in the tropics, mainly in mountainous regions rather than in lowlands, mainly in winter rather than in summer, and mainly at night rather than in daytime. The warming is real, but it is mostly making cold places warmer rather than making hot places hotter. To represent this* ***local warming*** *by a global average is misleading."*

Here is a nice actual map of local warming:



https://twitter.com/hausfath/status/1171820725390299136

Note the darker red up by the Arctic. Note also that spatial non-uniformities in temperature, whether vertically between surface and atmosphere, or between locations on the globe, are a major driver of weather phenomena, with for instance storm cyclones or large-scale air or ocean current flows being loosely analogous to the mechanical motion of a wheel driven by a heat engine. Michael Mann's online course has a nice lecture on a simple model of the climate system that includes a latitude axis.

The need for detailed and complex simulations

In any case, climate dynamics is clearly complicated. The complex spatial and feedback phenomena require global climate simulations to model, which need to be informed by measurements that are not yet 100% complete, some of which require observations over long timescales. Certain basic parameters that influence models, like the far-IR surface emissivity, are only beginning to be measured.

We've wandered far from our initial hope of doing *simple back of the envelope calculations* on the climate system, and have come all the way around to the perspective that **comprehensive, spatially fine grained, temporally long-term direct measurements of many key variables are needed** (and fortunately, many are ongoing), before this will be as predictable as one would like.

Integrating new mechanistic measurements and detailed calculations into efficiently computable global climate models is itself a challenge as well. A new project is making an improved climate model using the most modern data technologies, and there is a push for a CERN for climate modeling. Among current challenges against which modelers test and refine their simulations is the ability to accurately model the ENSO phenomenon, i.e., El Niño (meanwhile, a deep learning based statistical model has recently shown the ability to predict El Niño 1.5 years in advance).

Let's try to ballpark the feedback effects anyway!

This complexity has fortunately not deterred scientists from offering some coherent basic explanations, and building on these to make rough estimates of the "**climate sensitivity**"

to CO2: acs.org/content/acs/en/climatescience/atmosphericwarming/climatsensitivity.html

One approach for ball-parking the feedback effects, outside of complex many-parameter simulations, is to look at the past.

We want to understand how to relate the amount of warming we've seen so far to the amount of CO2 added **since pre-industrial times, when the CO2 concentration was around 278 parts per million** (ppm). Today, the CO2 concentration is around **408.02 ppm**.

(1 part per million of atmospheric CO2 is equivalent to 2.13 Gigatonnes of Carbon (not of CO2). So a pre-industrial level of 280 ppm is ~600 gigatonnes of Carbon, whereas our current level corresponds to 869.04 gigatonnes of Carbon, and doubling the preindustrial CO2 in the atmosphere, or 278 * 2 ~ 560 ppm, corresponds to about 1193 gigatonnes of Carbon. Note that GigaTonnes in the atmosphere doesn't translate directly into GigaTonnes emitted due to absorption of perhaps 1/3 or so by the oceans and perhaps another 1/3 or so by plants.)

Without any feedback:

If we *just* consider the direct absorption of the outgoing infrared by CO2 itself, one can get a formula like:

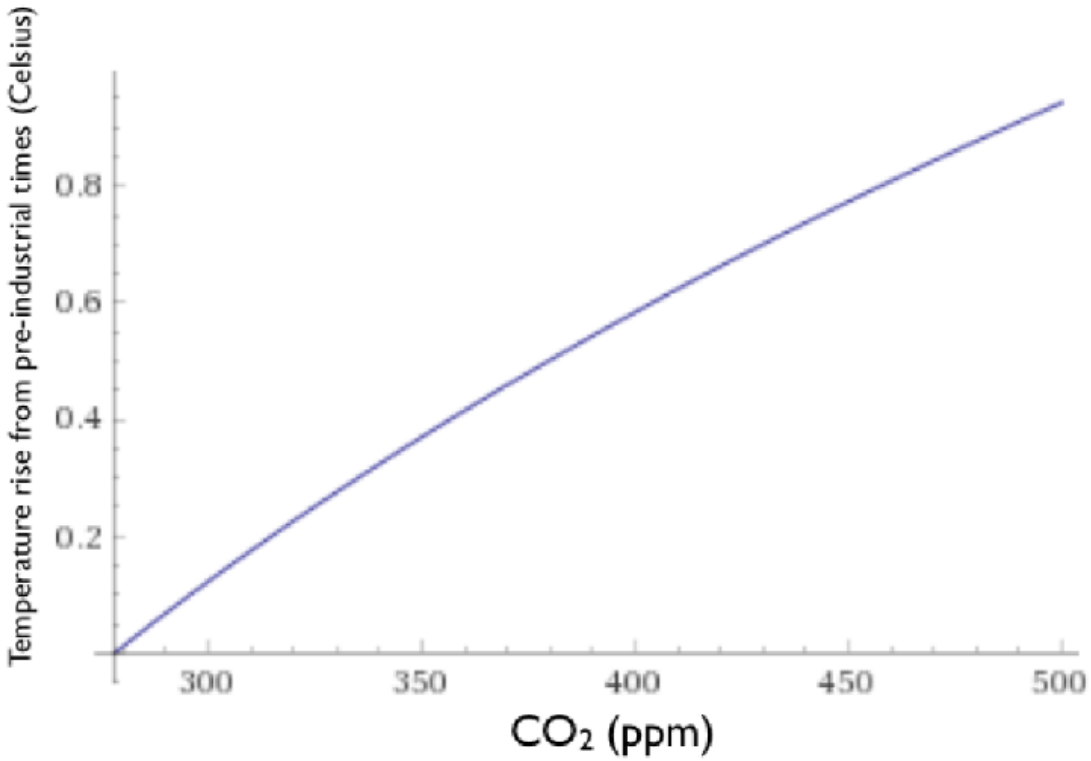
```
temperature_increase_from_preindustrial_levels

= [0.3 Kelvin per (Watt per meter squared)]

* (5.35 Watt per meter squared)

* ln([CO2 concentration in ppm]/278)
```

where ln(x) is the natural logarithm of x. Plugging this right into an online plotter, this gives the following curve (but see below, this is NOT the right answer):



(In addition to not including any feedback, this is also neglecting other greenhouse gasses than CO2.)

This formula comes in two parts. First, from a simple calculation using the radiation balance physics we've discussed already (specifically, linearizing a perturbation around the equilibrium we found above, see the sidebar at the ACS website for a derivation using the Stefan-Boltzmann formulas we used above), the change in temperature for a given amount of radiative forcing is

$$\Delta T \approx T_p * \Delta F / [4 (1 - A) S]$$

where ΔF is the radiative forcing from additional greenhouse gasses relative to those in pre-industrial times, T_p is the planetary surface temperature prior to additional greenhouse gasses being added ~ 288 Kelvin, A ~ 0.3 is the albedo and S is the average solar flux incoming, namely 1/4 of our 1376W/m^2 from above. If we this plug this in we get the 0.3

Kelvin per Watt per $meter^2$ term from the above formula.

Second, we have to get the radiative forcing ΔF from the CO2 concentration:
acs.org/content/acs/en/climatescience/atmosphericwarming/radiativeforcing.html
The 5.35 Watt per meter squared and the logarithmic dependence comes from the IPCC form used here.

Interestingly, there is a logarithmic dependence of ΔF on CO2 concentration. The fact that it is logarithmic is predicted from complex global simulations, but the basic physical reasons do not seem to be completely trivial:

- agupubs.onlinelibrary.wiley.com/doi/10.1029/98GL01908
  lam.mycpanel.princeton.edu/documents/LamAug07bs.pdf
- clivebest.com/blog/?p=4697

- scienceofdoom.com/2011/01/30/understanding-atmospheric-radiation-and-the-%E2%80%9Cgreenhouse%E2%80%9D-effect-%E2%80%93-part-three/
- http://www.realclimate.org/index.php/archives/2007/08/the-co2-problem-in-6-easy-steps/

One issue is that the absorption of infrared radiation by CO2 is associated with **broad spectral bands and sidebands**, and at the peaks of those absorption bands, absorption is **saturated** from the beginning — i.e., essentially every outgoing infrared photon from the surface with a wavelength right in the center one of these peaks will definitely be absorbed by CO2 long before it exits the atmosphere, even at pre-industrial CO2 concentrations — while regions near but not quite at the peaks may become saturated with increasing CO2 concentration.

Another key issue has to do with the ***height* in the atmosphere from which successful radiation out to space occurs**, and hence with the *temperature* at which it occurs — given the Stefan-Boltzmann T^4 dependence of the emitted radiation, and the fact that **temperature drops off sharply with height** into the atmosphere, *the rate of radiation will depend strongly on the height of the source in the atmosphere*.

If newly added CO2 effectively absorbs and emits IR radiation from **higher** in the atmosphere, its contribution to the outflux term of the radiative balance will be smaller, **warming** the planet, but this dependence **needn't be linear** on the CO2 concentration and in fact can be logarithmic. A bit more precisely, adding more CO2 to the atmosphere causes the *height* at which successful emission into space occurs to move higher, and hence colder, and hence *less*. It intuitively does, which this wonderful paper explains from the perspective of *convection* in the atmosphere: "the planetary energy balance is not purely radiative but is mediated by convection, with water vapor as the key middleman".

Plugging in our current 408 ppm to the above logarithmic formula, we get 0.6K of warming from pre-industrial times, versus around 1K observed in actual reality. A doubling of pre-industrial CO2 would then lead to about 1.1K of warming in this no-feedback toy model — not so bad. Alas, this is wrong — remember that we've left out the feedback.

With a historical estimate of feedback:

However, importantly, as the ACS website emphasizes, the formula we just used, which doesn't take into account any kind of water vapor or biological feedbacks, for example, not to mention many others, **under-estimates certain longer-timescale historical data**, taken from a record that spans over tens of thousands of years, by about a factor of about 4.

As the ACS website puts it:

> *"our calculated temperature change, that includes only the radiative forcing from increases in greenhouse gas concentrations, accounts for 20-25% of this observed temperature increase. This result implies a climate sensitivity factor [the term out front of the logarithmic expression for temperature change versus CO2 concentration]* ***perhaps four to five times greater, ~1.3 K·(W·m–2)–1, than obtained by simply balancing the radiative forcing of the greenhouse gases****. The analysis based only on greenhouse gas forcing has not accounted for feedbacks in the planetary system triggered by increasing temperature, including changes in the structure of the atmosphere."*

So what happens if we put in the missing factor of 4? Then the formula will be:

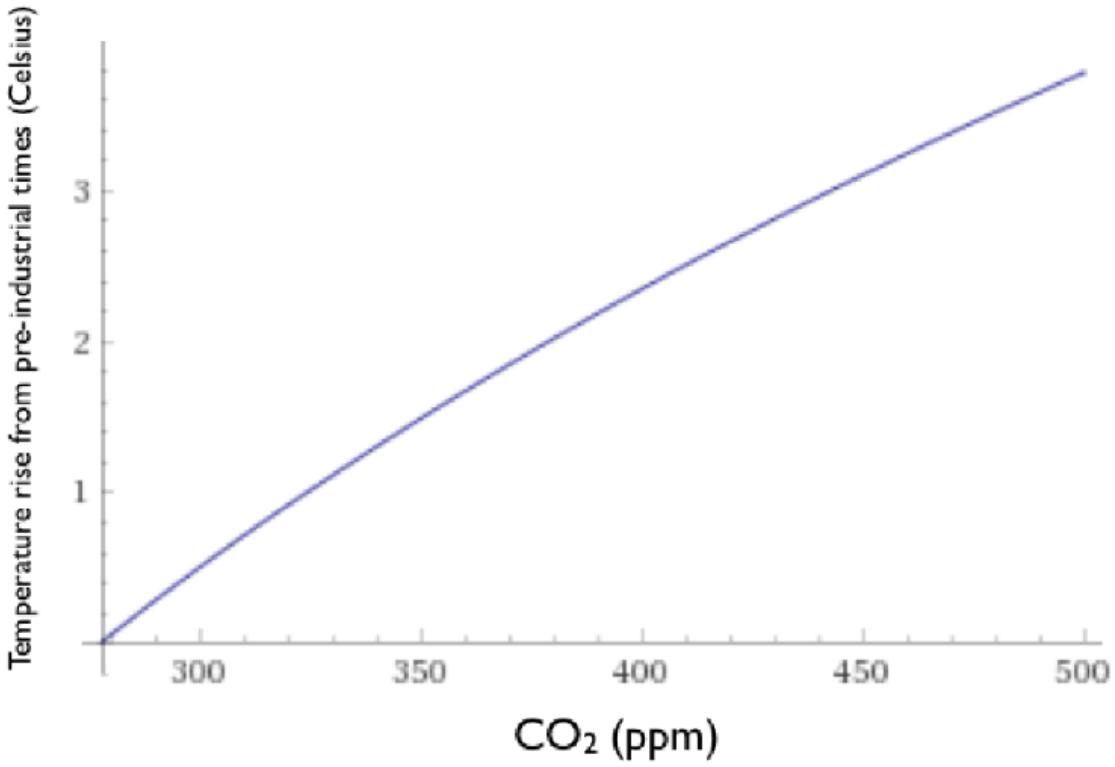
```
temperature_increase_from_preindustrial_levels

= 4 * [0.3 Kelvin per (Watt per meter squared)]

* (5.35 Watt per meter squared)
```

```
* ln([CO2 concentration in ppm]/278)
```

This looks like,



Then, even at our *current* ~408 ppm CO2 concentration, we would expect 2.4K of warming to *ultimately* (once the slow feedbacks catch up) occur, relative to pre-industrial times, if the missing factor of 4 is needed.

Now, this is much more warming than has occurred *so far*. What to make of that? Well, as mentioned above, the water-based feedback phenomena **can take decades to manifest** (given for instance the large heat capacity of the oceans), so perhaps we should not be surprised that *the temperature of the Earth has not yet caught up with where it is going to be heading over the coming decades*, even for a hypothetical *fixed* amount of CO2 from now on, according to this formula.

(Interestingly, researchers have now measured, using a sophisticated tool called the Atmospheric Emitted Radiance Interferometer, the raw CO2 forcing over "clear sky", i.e., *without* much of the long term or cloud related water vapor feedbacks, and found a CO2 forcing of around 1.82 W/meter^2, which **accords decently well** with what we get when we plug around 400 ppm CO2 into (5.35 Watt per meter squared) * ln([CO2]/278), namely (5.35 Watt per meter squared) * ln(400/278) = 1.95 W/meter^2. Unfortunately this can't tell us whether the "missing factor of 4" is indeed 4 or some other number — this is just the forcing from CO2 without water vapor effects.)

If true, i.e., if the (for instance) water feedbacks are indeed slow and positive to the point where the "missing factor of 4" is in fact needed to calculate the ultimate temperature impact, then even with the amount of CO2 that we've *already* emitted things are **not good**. Note that if CO2 concentration were (sometime later this century) to double from pre-industrial levels, in a model including this additional factor of 4, we'd have an average surface warming of about 4.5K above pre-industrial levels. That would be **really not good**.

A recent paper tried to pin down the magnitude of the water feedback based on more recent, rather than long-term historical, observations, including measurements of the water vapor concentration itself:

> *"the availability of atmospheric profiles of water vapor and temperature from the Atmospheric Infrared Sounder (AIRS) has made direct estimates of the water vapor feedback possible". They concluded that they could not perfectly estimate the vapor feedback without observations over a longer timescale than the 7 year long dataset they used, perhaps closer to 25 years: "The errors associated with this calculation, associated primarily with the shortness of our observational time series, suggest that the long-term water vapor feedback lies between 1.9 and 2.8 Wm−2K−1. The source of error from the relatively short time period is about three times as large as other error sources.... Thus,* ***our results strengthen the case for a significantly positive feedback from water vapor changes in the climate system, which, acting alone, would double the magnitude of any warming forced by increasing greenhouse gas concentrations****."*

So it sounds like they come out with roughly a factor of 2, rather than a factor of 4, for the effect of the water vapor feedback itself. Still not fantastic.

Potential tipping points

Alas, fitting historical or even quite recent data for the "missing feedback factor" isn't really fully predictive either, even if we had enough of that data, due to abrupt phase transitions or tipping points which can occur at different temperatures. For example, rapid one-time arctic methane release from the permafrost, or ocean albedo decrease brought on by sea ice melting. Some have argued that a nominal 2C target, given comparatively fast feedbacks like water vapor, actually corresponds to a nearly inexorable push to 3C or higher long-term when slower feedbacks are taken into account.

To be sure, many seem to disagree that the relevant tipping points are likely to be around 2C, as opposed to higher, as explained here. Here is a nice thread about the significance or lack thereof of current targets like 2C or 1.5C from a tipping point perspective. Also, many tipping points would take decades or centuries to manifest their effects, giving time for response, and in some cases in principle for reversal by bringing temperature back down through negative emissions.

But on the more worrisome side, I have seen plenty of papers and analyses that do point to the potential for major tipping points / bi-stabilities around 2C, or sometimes even less. For example, this paper uses 2C as a rough estimate of the temperature that you don't want to exceed for risk of hitting a potential *cascade* of major tipping points. Likewise this analysis points to potential tipping points in arctic permafrost methane release at even lower temperatures. (Also, we should probably worry about nonlinear effects that would depend on the *rate* of change of climate variables, for instance if ecologies can adapt to a given change over long timescales but struggle to do so over short timescales, and about the fact that some tipping point effects may not depend on average values of variables but on their peak values or variances.)

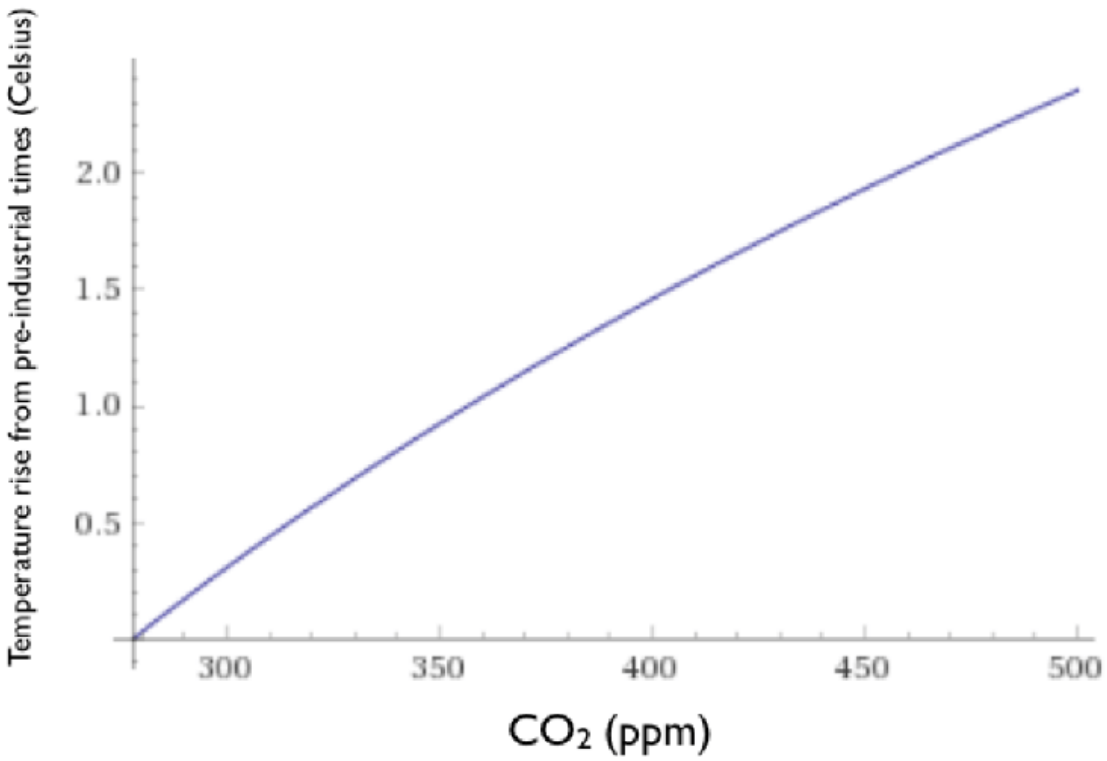
(As a side note, separate from "tipping points" in the climate system as a whole, you may have heard about dynamical chaos in the weather. These are two totally different issues, both related to nonlinear dynamics. Chaotic dynamics governing the specifics of weather in any particular "run" of the climate system, or particular run of a model thereof, is totally consistent with predictable and bounded trends in *average* weather, i.e., climate. This is beautifully explained here in reference to the canonical introductory example of chaos, the Lorentz attractor — in short, their analogy is that if *weather* is like any particular *trajectory* of the Lorentz system, unpredictable beyond a few weeks without exponentially more precise measurement of initial conditions, then *climate* is like the *overall shape and size* of the Lorentz attractor as a whole, on which all those trajectories must lie. By averaging enough

trajectories, you can start to understand aspects of that overall shape, which is what climate simulations do.)

Thus, we can start to see a *very rough qualitative picture* of the range of uncertainties. Nobody can *perfectly* pin down the exact climate sensitivity at the moment. But it is also now hopefully clear roughly where the estimates we see are coming from, and why they make sense given what is known.

A middle of the road climate sensitivity

For comparison, using sophisticated climate models, plus models of the human activity, IPCC projects 2-4K of warming by 2100. A classic meme is **450 ppm ~ 2 degrees C**, whereas the scaled logarithmic curve we have been discussing gives this equivalence if the "factor of 4" for feedback becomes a factor of **around 2.5**, which looks like



(Note on the status of this type of analysis: There is no reason that the curve *with* feedback has to be a simple *multiple* of the curve *without* feedback, i.e., that there is any simple "missing factor" at all, as opposed to a totally different shape of the dependency. Indeed, in scenarios with large abrupt tipping points this would certainly not be the case. You do, however, get something similar to this description when you linearize simple ordinary differential equation models for the planetary energy balance, by assuming a small perturbation around a given temperature — this is treated here. The site just linked arrives at a temperature perturbation deltaT ~ -R/lambda, where R is the radiative forcing and lambda is a feedback factor that arises from the linearization. In their analysis, they get lambda = −3.3 W m^-2 K^-1 due to just the direct effects of the CO2 forcing and blackbody radiation (called Planck forcing) and lambda = -1.3 W m^-2 K^-1 in a middle of the road IPCC climate sensitivity estimate: I think it is not coincidental here that -3.3/-1.3 ~ 2.5, which is the same as our "missing feedback factor" for the IPCC climate sensitivity estimate just above. In any case, the analysis here gives a more rigorous definition of feedback factors in linearized models and how to break them down into components due to different causes.)

We can also try to crudely translate this into the TCRE measure used in the IPCC reports. For us as back of the envelopers, I think this basically treats the logarithmic curve as linear over the range of interest, which you can see by looking at the curves above is not so far from the truth. If we use 450 ppm for 2C, then our TCRE is 2.7C/1000PgC, whereas if we use 480 ppm for 2C it is 2.34C/1000PgC, and if we use 530 ppm for 2C it is 1.9C/1000PgC. The IPCC states (based on statistically aggregating results of complex simulations and so forth) a likely range for the TCRE of 0.8 to 2.5C/1000PgC with a best observational estimate of 1.35C/1000PgC, suggesting that our feedback factor of 2.5 may be a bit high and a feedback factor of 2 may be more appropriate as a middle of the range estimate.

Implications of the middle of the road climate sensitivity for carbon budgets

What would this particular curve (used for illustration only, since any real treatment has to be probabilistic, and this model is extremely crude, see the IPCC reports for better), with the "feedback factor" of 2.5, imply for the hypothetical scenario where we abruptly *ceased* emissions quite soon?

Per the curve, we'd still potentially be looking at around 1.5C of ultimate warming (due to feedbacks), with roughly our *current* atmospheric CO2 levels kept *constant*.

With the feedback factor of 2 instead of 2.5, we'd cross 1.5C at around 440 ppm, and our remaining carbon budget for 1.5C at the time of writing would therefore be around 2*(440-405 ppm)*7.81=547GtCO2 emitted.

For comparison, one figure I've seen (see slide 15 there) gives around 0.1C per decade of additional warming even under year 2000 atmospheric *CO2* ***levels*** kept **constant**, although another analysis suggests roughly constant temperature (little further warming) if carbon **emissions** were to completely and abruptly **cease** — these are not inconsistent, since if emissions somehow *completely* ceased, natural carbon sinks could start to kick in to *slowly* decrease atmospheric CO2, which is not the same as holding CO2 *concentrations* constant. See more good discussion here and here. It is not obvious to me a priori that some fixed amount of CO2 added to the atmosphere would be naturally removed, on net, at all, even on long timescales... a carbon cycle could in principle keep atmospheric CO2 at a stable level that simply happens to be higher than the pre-industrial concentration. But according to a nice discussion in the introduction to the NAS report on negative emissions, the sinks are real and active, and indeed once human emissions fall *low enough*, not even necessarily to zero, the sinks can bring the rate of change of atmospheric CO2 concentration to net negative:

> *"...atmospheric concentration declines... despite net anthropogenic emissions of 164 Gt CO2 during the same period (fossil and land use minus NET), because the land and ocean carbon sinks take up 196+164 Gt CO2... the sinks are expected to persist for more than a century of declining CO2 because of the continued disequilibrium uptake by the long-lived carbon pools in the ocean and terrestrial biosphere. For example, to reduce atmospheric CO2 from 450 to 400 ppm, it would not be necessary to create net negative anthropogenic emissions equal to the net positive historical emissions that caused the concentration to increase from 400 to 450 ppm. The persistent disequilibrium uptake by the land and ocean carbon sinks would allow for achievement of this reduction even with net positive anthropogenic emissions during the 50 ppm decline"*

Anyway, loosely speaking (see here for a much better analysis), from what I've understood, in the absence of large-scale deployment of negative emissions technology, it seems *unlikely* that we have more than 10 years of emissions *at current levels* while still having a *high likelihood* of not exceeding a 1.5C peak warming, given what we know now. That's ~400 GigaTonnes CO2 for the 10 years, versus a budget of roughly zero for a feedback factor of 2.5 and roughly 550GtCO2 for a feedback factor of 2 as estimated above. There is a *chance* we've already committed to 1.5C peak warming, in the absence of large-scale negative emissions, regardless of what else we do.

Interestingly, even some of the scenarios that target 1.5C set that as an ultimate target for the end of the century, and actually slightly exceed that target between now and then, relying on negative emissions to bring it down afterwards. This recent Nature paper explains clearly why and how not to think that way: we need to make our near-term action more stringent to constrain the peak, and the timing of the peak, not just a long-term target.

With all that in mind, here are two snippets from the IPCC report on 1.5C (written over a year ago so the carbon budget is roughly 40-80 GtCO2 less now). The left hand snippet expresses the level of uncertainty in the remaining carbon budget for 1.5C, and the right snippet mentions the need for negative emissions in real-world scenarios where we stabilize to 1.5C. Their 420-580 GtCO2 budget from a year or so ago accords decently well with what we got (zero to 550 GtCO2) from the feedback factor of 2-2.5 case above.

Cumulative CO_2 emissions are kept within a budget by reducing global annual CO_2 emissions to net zero. This assessment suggests a remaining budget of about 420 $GtCO_2$ for a two-thirds chance of limiting warming to 1.5°C, and of about 580 $GtCO_2$ for an even chance (*medium confidence*). The remaining carbon budget is defined here as cumulative CO_2 emissions from the start of 2018 until the time of net zero global emissions for global warming defined as a change in global near-surface air temperatures. Remaining budgets applicable to 2100 would be approximately 100 $GtCO_2$ lower than this to account for permafrost thawing and potential methane release from wetlands in the future, and more thereafter. These estimates come with an additional geophysical uncertainty of at least ±400 $GtCO_2$, related to non-CO_2 response and TCRE distribution. Uncertainties in the level of historic warming contribute ±250 $GtCO_2$. In addition, these estimates can vary by ±250 $GtCO_2$ depending on non-CO_2 mitigation strategies as found in available pathways. {2.2.2, 2.6.1}

The Role of Carbon Dioxide Removal (CDR)

All analysed pathways limiting warming to 1.5°C with no or limited overshoot use CDR to some extent to neutralize emissions from sources for which no mitigation measures have been identified and, in most cases, also to achieve net negative emissions to return global warming to 1.5°C following a peak (*high confidence*). The longer the delay in reducing CO_2 emissions towards zero, the larger the likelihood of exceeding 1.5°C, and the heavier the implied reliance on net negative emissions after mid-century to return warming to 1.5°C (*high confidence*). The faster reduction of net CO_2 emissions in 1.5°C compared to 2°C pathways is predominantly achieved by measures that result in less CO_2 being produced and emitted, and only to a smaller degree through additional CDR. Limitations on the speed, scale and societal acceptability of CDR deployment also limit the conceivable extent of temperature overshoot. Limits to our understanding of how the carbon cycle responds to net negative emissions increase the uncertainty about the effectiveness of CDR to decline temperatures after a peak. {2.2, 2.3, 2.6, 4.3.7}

https://www.ipcc.ch/site/assets/uploads/sites/2/2019/02/SR15_Chapter2_Low_Res.pdf

https://www.ipcc.ch/site/assets/uploads/sites/2/2019/02/SR15_Chapter2_Low_Res.pdf

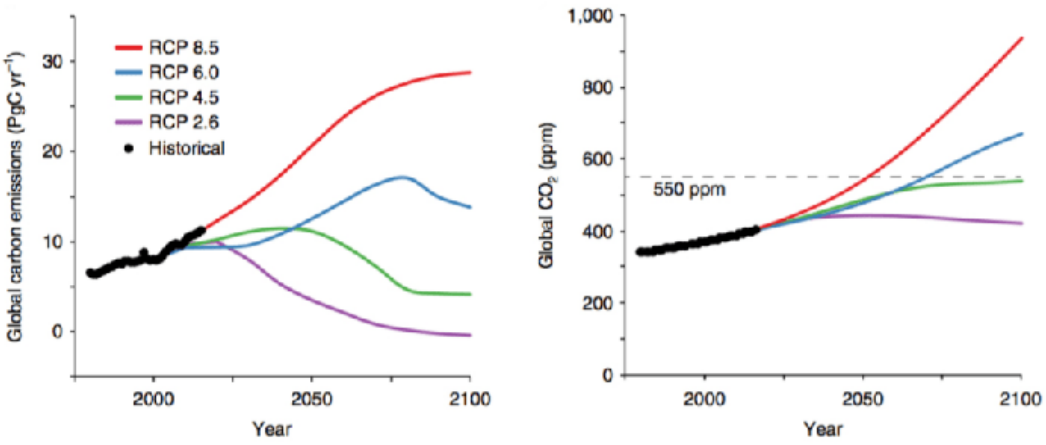
We can also ballpark (again, subject to all the caveats about the notion of carbon budgets here, not to mention our crappiest of models) a remaining carbon budget for 2C warming using the 450 ppm ~ 2C idea. 450 ppm ~ 3514 GigaTonnes of CO2 in the atmosphere, versus ~3163 (405 ppm) up there now, so we would be allowed to add around 300 or 350 more. So if the oceans and land plants absorb 1/2 of what we emit (bad for ocean acidification), the maximum carbon budget for 2C in this very particular scenario is around 2*(3514-3163) GigaTonnes CO2 = 702 GigaTonnes of emitted CO2 if there are no negative emissions. At the time of writing, meanwhile, this calculator from the Guardian, gives 687 GigaTonnes, or 17 years at our current emissions rate. Now, the carbon budgets I'm seeing in the IPCC reports for 2C seem a bit higher, in the range of 1200-2000 GigaTonnes, which we can get if we use 480 ppm as our target for 2C instead of 450 ppm: 2*(480-405 ppm)*7.81 GtCO2/ppm = 1171 GtCO2, or if we use 530 ppm as our 2C target we get 2*(530-405)*7.81=1953 GtCO2.

Impacts

What might this translate to in terms of impacts? Let's consider just one: sea level rise.

Some baseline scenarios for sea level rise

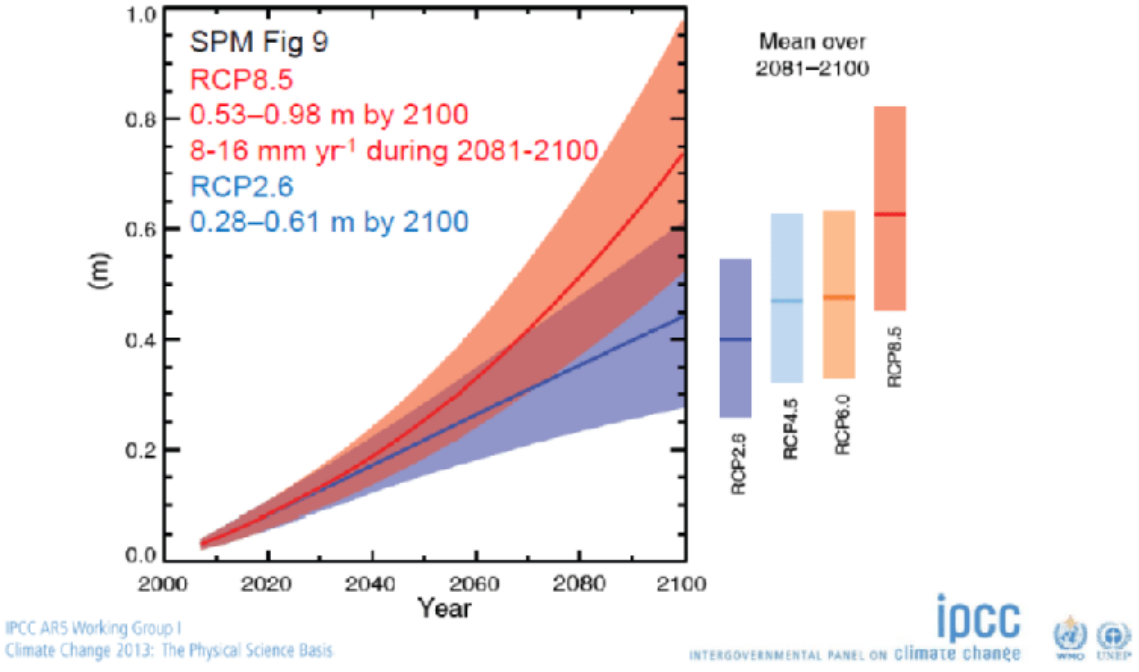
Here are some emissions scenarios, and then the corresponding projected mean sea level changes from IPCC:



Left: Global carbon emissions (black) and projected emissions under RCP2.6 (purple), RCP4.5 (green), RCP6.0 (blue) and RCP8.5 (red) from 1980-2100. Right: Global CO2 ppm from 1980-2100, with dashed line indicating point at reach concentrations reach 550ppm. Source: Smith & Myers (2018)

https://www.researchgate.net/figure/Historical-trends-in-CO2-emissions-and-atmospheric-concentrations-compared-with-model_fig1_327269367

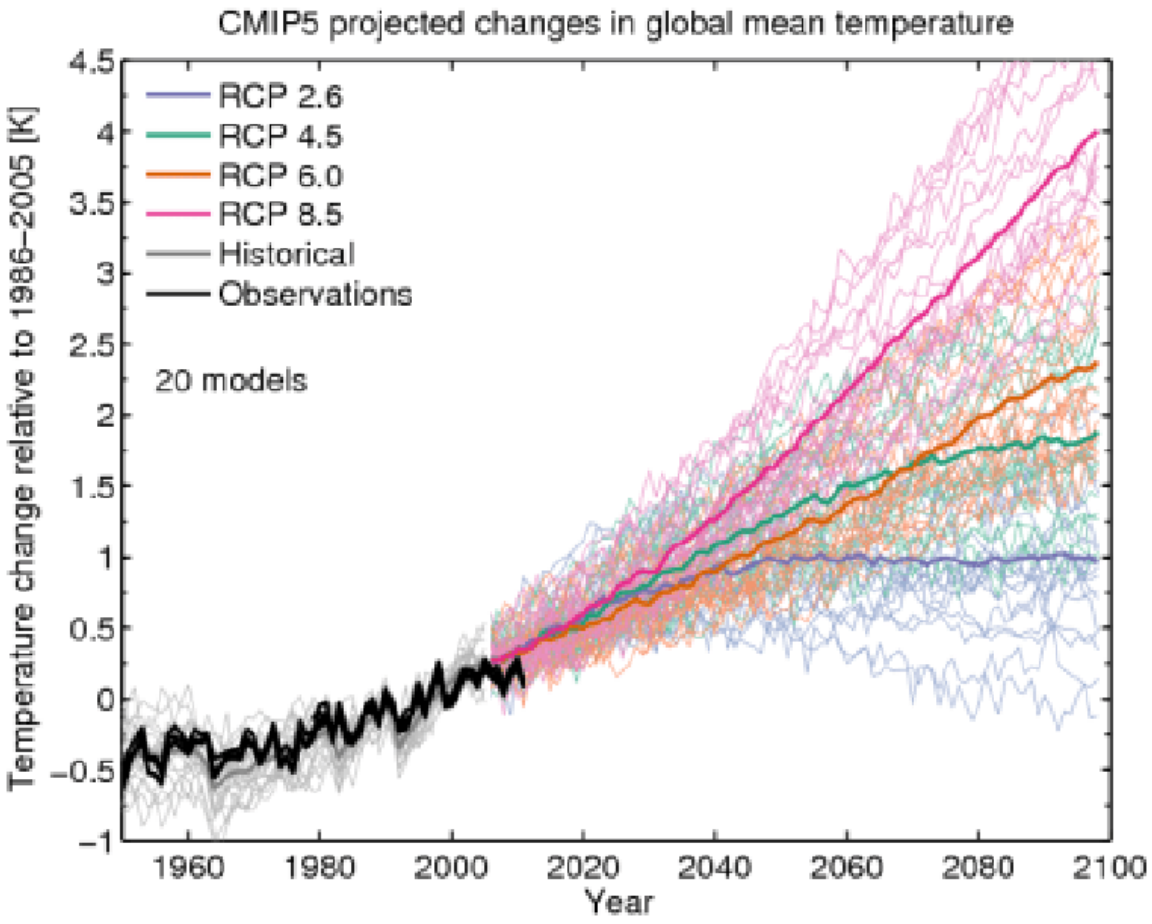
https://www.researchgate.net/figure/Historical-trends-in-CO2-emissions-and-atmospheric-concentrations-compared-with-model_fig1_327269367



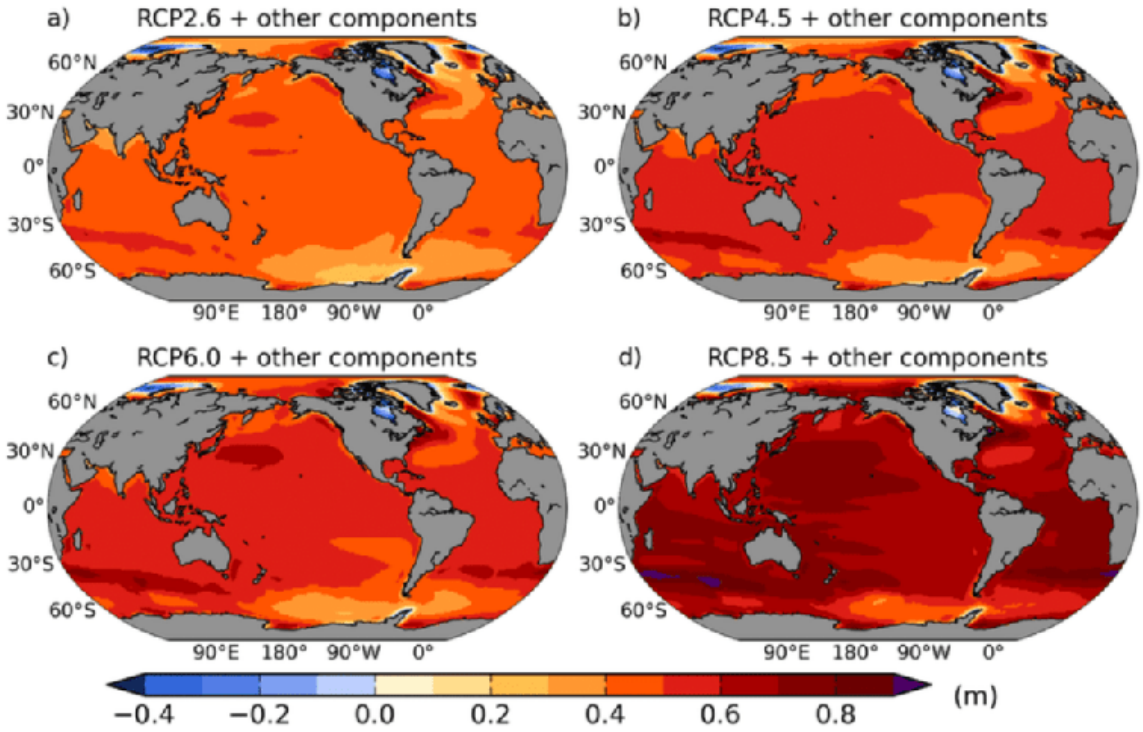
https://www.metlink.org/climate/ipcc-2013-figures/

https://www.metlink.org/climate/ipcc-2013-figures/

Here is temperature in those same scenarios:



Here is a map of the expected sea level rise globally in those same scenarios:



20: Ensemble mean net regional sea level change (m) evaluated from 21 CMIP5 models for the RCP scenarios (a) 2.6, (b) 4.5, (c) 6.0 and (d) 8.5 between 1986-2005 and 2081-2100. Each map includes effects of atmospheric loading, plus land-ice, GIA and terrestrial water sources.

https://www.researchgate.net/figure/Ensemble-mean-net-regional-sea-level-change-m-evaluated-from-21-CMIP5-models-for-the_fig6_284695835

https://www.researchgate.net/figure/Ensemble-mean-net-regional-sea-level-change-m-evaluated-from-21-CMIP5-models-for-the_fig6_284695835d

According to Michael Mann's course, at 1 meter global average sea level rise, 145 million people would be directly affected, and there would be on the order of a trillion dollars of damage. Note that the rise at certain coastal areas will be higher than the global average. This paper provides some specifics:

> *"By 2040, with a 2 °C warming under the RCP8.5 scenario, more than 90% of coastal areas will experience sea level rise exceeding the global estimate of 0.2 m, with up to 0.4 m expected along the Atlantic coast of North America and Norway. With a 5 °C rise by 2100, sea level will rise rapidly, reaching 0.9 m (median), and 80% of the coastline will exceed the global sea level rise at the 95th percentile upper limit of 1.8 m. Under RCP8.5, by 2100, New York may expect rises of 1.09 m, Guangzhou may expect rises of 0.91 m, and Lagos may expect rises of 0.90 m, with the 95th percentile upper limit of 2.24 m, 1.93 m, and 1.92 m, respectively."*

Implications of potential tipping points in ice melt

The above plots of mean/expected sea level rise do **not** strongly reflect specific, apparently-much-less-likely scenarios in which we would see catastrophic tipping-point collapse of the Arctic or Antarctic ice.

To be up front, I have limited understanding of how likely this is to happen and what the real implications would be. But to try to understand this, it looks helpful to separate major sources of sea level rise due to ice melt into at least four groups:

- Sea ice (Arctic and Antarctic)
- Greenland ice sheet
- West Antarctic ice sheet
- East Antarctic ice sheet

We are already seeing significant loss of summer sea ice in the Arctic. In Antarctica, the sea ice normally doesn't survive in summer, but loss of the winter sea ice could significantly disrupt ocean circulation.

If Greenland melted in full, sea level would rise around 7 meters.

If *West* Antarctica melted in full, we'd be seeing several meters of sea level rise.

If all of Antarctica, including the comparatively-likely-to-be-stable East Antarctica, melted we'd be seeing perhaps > 60 meters of sea level rise.

(Sea level also rises with rising temperature just due to thermal expansion of water, as opposed to ice melt.)

Ideally ice melt would go slowly, e.g., a century or more, giving some time to respond with things like negative emissions technologies (next post) and perhaps some temperature stabilization via solar geo-engineering (next next post). But from my understanding, it is at least *conceivable* that it could happen in just a couple of decades, for instance if a series of major tipping points in positive feedbacks are crossed, e.g., methane release en masse from the permafrost. For comparison, the city of Manhattan is 10 meters above sea level.

Now, a question for each of these is whether they can be *reversible*, if we brought back the temperature via removal of CO2 or via solar radiation management.

We know that there is a weak form of what could be called "effective irreversibility by emissions reduction" in the sense that the climate has inertia, so that if we simply greatly reduced emissions, it might be too late — a self-reinforcing process of warming and sea ice melt could have already set in due to positive feedbacks. For instance, as ice melts, ocean albedo lowers, melting more ice. Moreover, we've already discussed that there are water vapor feedbacks that take time to warm the planet fully once CO2 has been added to the atmosphere. Thus, we can effectively lock in ongoing future sea level rise with actions that are happening or already have happened. This seems to be a concept used frequently in popular discussions of this topic, i.e., people ask "is it already too late to stop sea level rise by limiting emissions".

But there is also a stronger notion of irreversibility — if we used negative emissions or solar radiation management to restore current temperatures, would we also restore current ice formations?

In one 2012 modeling paper called "How reversible is sea ice loss", the authors concluded that: "Against global mean temperature, Arctic sea ice area is reversible, while the Antarctic sea ice shows some asymmetric behaviour – its rate of change slower, with falling temperatures, than its rate of change with rising temperatures. However, we show that the asymmetric behaviour is driven by hemispherical differences in temperature change between transient and stabilisation periods. ***We find no irreversible behaviour in the sea ice cover***". In other words, if you can bring the temperature back down, you can bring back the sea ice, in this study, with some delay depending on how exactly you bring down the temperature. But this is just one modeling study.

Another paper called "The reversibility of sea ice loss in a state-of-the-art climate model", concludes similarly for the sea ice as such:

> *"The lack of evidence for critical sea ice thresholds within a state-of-the-art GCM implies that future sea ice loss will occur only insofar as global warming continues, and may be fully reversible. This is ultimately an encouraging conclusion; although some future warming is inevitable [e.g., Armour and Roe, 2011], in the event that greenhouse gas emissions are reduced sufficiently for the climate to cool back to modern hemispheric-mean temperatures,* ***a sea ice cover similar to modern-day is expected to follow****."*

So maybe the sea ice cover, as such, is reversible with temperature — albeit with a *delay*: their notion of reversibility is defined in a simulation that unfolds over multiple centuries!

But, perhaps even more worryingly, this doesn't tell us anything about reversibility of Greenland or West or East Antarctica melt: this is ice that rests on land of some kind, as opposed to ice that arises from freezing in the open seas. In fact, the second reversibility study says: "our findings are expected to be most relevant to the assessment of sea ice thresholds under transient warming over the next few centuries **in the absence of substantial land ice sheet changes**".

A paper in PNAS provides a relatively simple model and states:

> *"We also discuss why, in contrast to Arctic summer sea ice, a tipping point is more likely to exist for the loss of the Greenland ice sheet and the West Antarctic ice sheet."*

They mention at least some evidence for irreversibility for Greenland land ice: "…there is some evidence from general circulation models that the removal of the Greenland ice sheet would be irreversible, even if climate were to return to preindustrial conditions (54), which increases the likelihood of a tipping point to exist. This finding, however, is disputed by others (55), indicating the as-of-yet very large uncertainty in the modeling of ice-sheet behavior in a warmer climate (56)." They also note that: "In this context, it should be noted that the local warming needed to slowly melt the Greenland ice sheet is estimated by some authors to be as low as 2.7° C warming above preindustrial temperatures (64, 65). This amount of local warming is likely to be reached for a global warming of less than 2° C (39)."

So we're at real risk of melting Greenland ice, and doing so *might* not be easily reversible with temperature after the fact. Its *loss* could probably be slowed or hopefully stopped by temperature restoration, especially by solar radiation management — a recent paper reports that "Irvine et al. (2009) conducted a study of the response of the Greenland Ice Sheet to a range of idealized and fixed scenarios of solar geoengineering deployment using the GLIMMER ice dynamics model driven by temperature and precipitation anomalies from a climate model and found that under an idealized scenario of quadrupled CO2 concentrations *solar geo-engineering could slow and even prevent the collapse of the ice sheet*."

West Antarctica looks, if anything, worse. Notably, there is at least the possibility of bistable dynamics or very strong hysteresis whereby Antarctic melting, once it got underway, could not be reversed even if we brought the temperature and atmospheric CO2 back to current levels. It is apparently only recently that quantitative modeling of the ice dynamics has made major strides.

In particular, the West Antarctic Ice Sheet is vulnerable to a so-called Marine Ice Sheet Instability, which some suggest may already be occurring, and which depends on the specific inward-sloping geometry of the land underlying that ice formation. It is possible that this dynamic is *already* starting to happen. (Not to be confused with the Marine Ice *Cliff* Instability, which isn't looking too bad.)

Conceivably, though not by any means with certainty from what I can tell, *even* solar radiation management, which would artificially bring the *temperature* back down by lowering solar influx through the atmosphere, might not be able to stop the *melting* when it is driven

by such unstable undersea processes: in the case of the Antarctic, one recent paper suggested that

> *"For Antarctica where ice discharge is the dominant loss mechanism, the most significant effect of climate change is to thin and weaken ice shelves which provide a buttressing effect, pushing back against the glaciers and slowing their flow into the ocean... whilst stratospheric aerosol geoengineering... could lower surface melt considerably it may have a limited ability to reduce ice shelf basal melt rates... there is the potential for runaway responses which would limit solar geoengineering's potential to slow or reverse this contribution to sea level rise".*

In other words, even if you could bring the temperature back to that of today (and I believe this could apply to restoring the CO2 level via negative emissions as well), you *might* not be able to stop the melting of West Antarctica.

All is not completely lost in that case, to be sure. There are at least some proposed local glacier management schemes that could directly try to limit the ice loss, including buttressing the ice sheet underwater, as well as spraying snow on top. Dyson somewhat fancifully suggested actively lowering sea levels by using a fleet of tethered kites or balloons to alter local weather patterns such that we would dump snow on comparatively stable *East* Antarctica, thus removing water from the oceans and actively lowering sea levels, irrespective of what caused those sea levels to rise in the first place. In short, given enough urgency and time humanity will likely find a way — and there are probably other ideas beyond Dyson's kites that could actively reduce sea levels and counteract sea level rise regardless of its exact cause.

But seriously, we want to avoid major tipping points in the climate system. If we cross major tipping points in sea or land ice melt, it will be very challenging to deal with at best.

None of this is to say exactly how likely or unlikely it is that such tipping points would be reached. I simply don't know, and I haven't found much online to ground an answer in a specific temperature or timeframe — let me know if you have something really good on this. It is *hopefully* very unlikely even a bit above 2C warming, and hopefully very slow as well, giving more time to respond and, to the extent possible, reverse the melting.

But there is some real possibility that self-accelerating processes of ice melt in Greenland and West Antarctica are already happening. Here is a recent update on Antarctica:

> *"Most of Antarctica has yet to see dramatic warming. However, the Antarctic Peninsula, which juts out into warmer waters north of Antarctica, has warmed 2.5 degrees Celsius (4.5 degrees Fahrenheit) since 1950. A large area of the West Antarctic Ice Sheet is also losing mass, probably because of warmer water deep in the ocean near the Antarctic coast. In East Antarctica, no clear trend has emerged, although some stations appear to be cooling slightly. Overall, scientists believe that Antarctica is starting to lose ice, but so far the process has not become as quick or as widespread as in Greenland."*
> ***nsidc.org/cryosphere/quickfacts/icesheets.html***

To summarize:

- Sea ice: hopefully fairly quickly reversible
- Greenland melt: unclear, maybe reversible by solar radiation management or aggressive negative emissions
- West Antarctica: may not be reversible at all except by things like active snow dumping via altering air flows, or by preventing the melt through things like direct buttressing of the ice sheet
- East Antarctica: let's hope it doesn't melt… fortunately it seems unlikely to do so

Of course there are many other potential impacts, e.g., droughts.

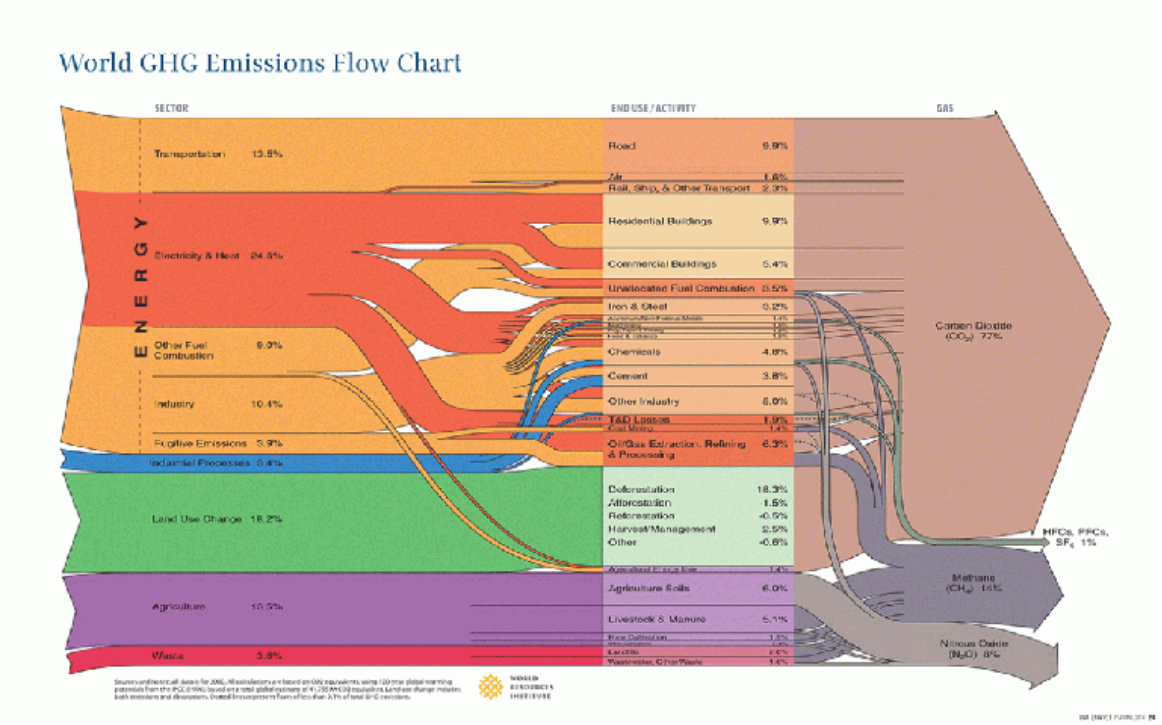
Further reading: IPCC technical summary, NAS report on climate, summary of IPCC report on the oceans.

Emissions

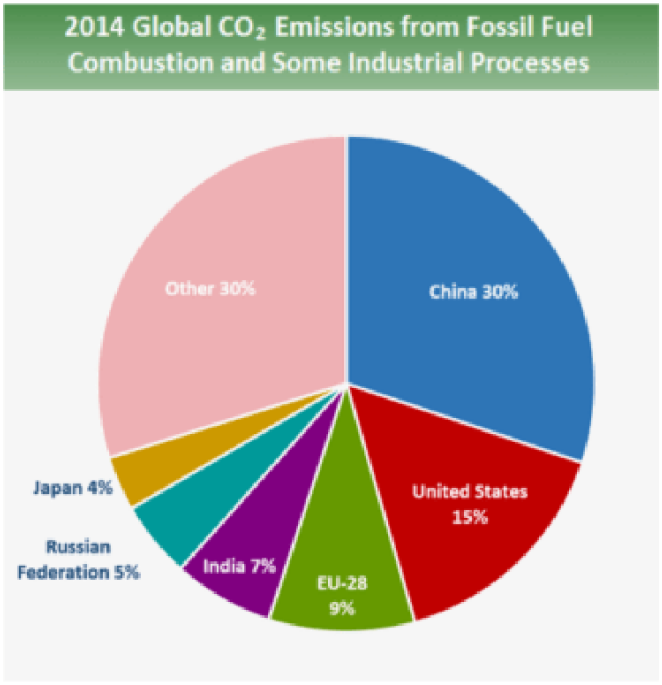
Note: This part starts to make a tiny bit of contact with human social, economic and cultural complexities. You can take it as just my still-evolving personal opinion, with my acknowledgedly-limited understanding of human society and how it might change.

Sectors driving emissions

Electricity per se is only ~25% of global greenhouse gas emissions. There are many gigatonnes that we have no large-scale alternative for today, like those associated with airplanes, steel, cement, petrochemicals, and seasonal energy storage. Here is a "Sankey Diagram" from the World Resources Institute:



Here is a breakdown of emissions globally:



Source: Boden, T.A., Marland, G., and Andres, R.J. (2017). National CO2 Emissions from Fossil-Fuel Burning, Cement Manufacture, and Gas Flaring: 1751-2014, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, doi 10.3334/CDIAC/00001_V2017.

https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#Sector

https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#Sector

Breaking down emissions in terms of sectors and global socioeconomics is sobering. Recent pledges have barely moved the needle. We need major changes in the structure of the industrial economy at large, globally. Rich countries have a key role in bringing down the cost of decarbonization for poorer countries, at scale, e.g., through developing and demonstrating better, cheaper technology.

A recent paper estimated that, to remain within a 1.5C target, not only must we curtail the building of new carbon-intensive infrastructure *quickly* — but also, a large amount of existing fossil fuel based infrastructure will need to be *retired early and replaced* with clean alternatives (assuming these exist in time, at feasible cost levels). Many factors — including the growth of China and India, the likely availability of large-scale deposits of new fossil fuels, as well as the large and diverse set of industries that emit greenhouse gasses, not to mention ongoing Western policy dysfunctionality — currently make this very difficult. On the plus side, Ramez Naam argues that building new renewables is becoming cheaper than operating existing coal. That's the electricity sector, but other sectors like industrial steel and cement production, land use, heating, agriculture and aviation fuel do not have an equivalent cost-effective fossil fuel replacement yet — Ramez's plan is very much worth reading in both regards, including for its suggestions for ARPA-style agencies for agriculture/food and industry/manufacturing.

Evaluating some popular ideas about emissions reduction

I found two examples interesting, from my brief attempt to gain some emissions numeracy:

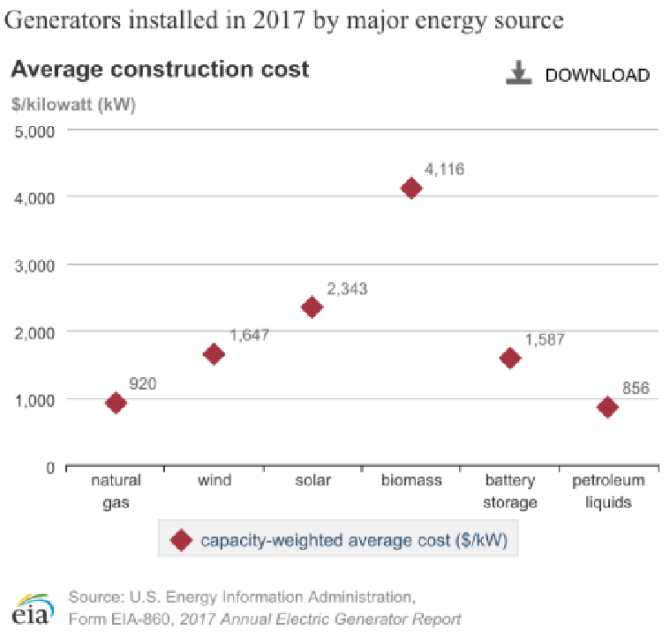
Example 1: Electric cars. We'll use a rough figure of 250,000 non-electric cars = 1 megaton CO2 emitted per year. Now, there are roughly a billion cars in the world, so 1e9/2.5e5 * 1 megatonne = 4 gigatonnes CO2 from cars, out of about 37 emitted total. About 11% cars by that estimate. Transportation overall is 28% of emissions in the US, 14% globally. If we electrified all cars instantly, the *direct* dent in the overall problem is thus fractional but significant. But electrification of cars has implications for other important aspects, like renewables storage (e.g., batteries) and the structure of the power grid, and for the price of oil. We haven't considered the full life-cycle effects of manufacturing and building electric versus conventional cars here, e.g., battery manufacturing, but it is clear that electric cars are very positive long term and have the potential to massively disrupt the electricity sector as well as the transportation fuel sector. Musk points out here that even *without* renewables, large-scale energy storage can decrease the need for coal or natural gas power plants by buffering over time. We'll discuss the role of driving down battery costs in the broader picture of renewable electricity below.

Example 2: Academic conference travel. Many of my colleagues are talking about reducing their plane flights to academic conferences in order to help fight climate change. My intuition was that this is not so highly leveraged, at least as individual action rather than a cultural signal of urgency to push large-scale action by big institutions and governments. After all, emissions from air travel are a small fraction of the total emissions, and the fraction of air travel due to the scientific community is small. (Fortunately there are other highly leveraged ways for my colleagues in neuroscience or AI to contribute outside the social/advocacy sphere.) But let's break down the issue more quantitatively. Say your flights to/from a far-away conference once yearly are on the order of 3 tonnes CO2 (this was for a round-trip London to San Diego flight), versus about 37 billion tonnes total CO2/year emitted globally. Say there are 2 million scientists traveling to conferences each year (the Society for Neuroscience meeting alone is 30k, but not all researchers travel to far-away conferences every year). That gives 6 million tonnes of CO2 from scientists traveling to conferences. Say it is 10-20 million tonnes of CO2 in reality, emitted yearly. That's a 0.03-0.06% reduction of global emissions if we replaced it with, say, video calls. The carbon offsets for this CO2 appear to be around $100M-$200M a year (on the order of $10/tonne CO2).

A bit about clean electricity, renewables, storage, nuclear, and the power grid

I won't talk *much* about clean energy generation here, instead pointing readers to David MacKay's amazing book and this review paper. But I will mention a few quick things about the scale of the problem, and about a few technologies.

According to this article, decarbonizing world energy consumption by 2050 would require building the equivalent of a nuclear power plant every 1.5 days, or 1500 wind turbines per day, from now till 2050, not to mention taking down old fossil fuel energy plants. Let's work this out ourselves, and estimate a ballpark capital cost for it. Let's use the figure of needing to decarbonize 16 TeraWatts (TW) or so, without introducing negative emissions technologies. A typical coal plant is, say, 600 MW or so. 16 TW / (600 MW) / (30 years *365 days per year) = equivalent of replacing 2.5 coal plants per day with non-fossil energy. Now let's consider the cost of constructing new energy infrastructure. We'll assume that constructing renewables or nuclear can be made cost-competitive with constructing fossil fuel plants. A new fossil fuel based plant costs say on the order of $1000/kW



So we're talking $1000/kW * 16 TW / (30 years) = $500 billion per year. That's half a percent of world GDP every year as a kind of lower-end estimate of up-front capital costs, barring much better/cheaper technology.

Nuclear

The details are continually vigorously debated, but aggressive development and deployment of improved nuclear power systems (including probably fusion, *long* term, which could be very compact, long-lasting, and safe, although it is unclear it can come "in time", from the perspective of significantly contributing to limiting the *peak* CO2 concentration in the atmosphere over the coming 2-3 decades — it is possible that it can, but a stretch and definitely unproven) seems like an important *component*, at least in some countries. It is worth reading David MacKay's blunt comments on the subject, as it applies to the UK specifically, just before his untimely death.

Some would propose to keep existing nuclear plants operating, but wait for the cost, speed of deployability, safety and performance of next-generation nuclear systems — like "small modular reactors" — to improve before deploying new nuclear. But the idea that existing nuclear is fundamentally too expensive in the current regulatory environment is perhaps at least debatable. Nuclear is the largest single source of energy in France with 58 reactors — although the just-linked post points out that the French nuclear industry is in financial trouble. In this blog post Chris Uhlik calculates what it would take to scale up nuclear technologies at *current cost levels*, and concludes the following

> **Result:** In this future, we need 7.7 kW per person, provided by $3/watt capitalized sources with 8% cost of capital and 35% surcharge for O&M. The cost of this infrastructure: $2,550/person/year or 5% of GDP.
>
> *Alternate assumptions*:
>
> - Chinese nuclear plant costs of $1.70/watt
> - Higher efficiency in an electric future were most processes take about 1/2 as much energy from electricity as they used to take from combustion. 1.3 kW from old electricity demands (unchanged) + 3.2 kW from new electricity demands (half of 6.4 kW). And fuels (where still needed) are produced using nuclear heat-driven synthesis approaches.
>
> *Alternative result*: $844/person/year or 2% of GDP.
>
> **Conclusion:** Saving the environment using nuclear power could be cheap and worth doing.

https://bravenewclimate.com/2011/01/21/the-cost-of-ending-global-warming-a-calculation/

https://bravenewclimate.com/2011/01/21/the-cost-of-ending-global-warming-a-calculation/

So that's about 5-10x as much as our 0.5% world GDP per year lower-end estimate above, based on the approximate cost of fossil plants, a cost level which new renewables appear to be rapidly approaching. Uhlik still calls it "cheap", and that doesn't seem unreasonable given the benefits.

Is Uhlik missing anything? Well, one key issue is the fuel. There are two issues within that: 1) availability, and 2) safe disposal. As far as availability, this review paper suggests that getting enough for *traditional* U235 reactors to completely power the planet for decades may be a stretch. They suggest either using breeder reactors, such as thorium reactors, or getting fusion to work, to get around this. You might thus not be surprised that Uhlik is involved in thorium reactor projects — indeed I think his post *should be understood in the context of scaling up thorium reactors*, not existing uranium light water reactors. Thorium is still at an early stage of development, although some companies are trying to scale up the existing experiments. For two contrary perspectives on thorium, see here and here. Beyond thorium per se, though, there are many other apparently promising types of molten salt reactors.

The waste disposal problem seems complex and depends on the reactor technology. Moreover, some novel reactors could be fueled on the waste of others, or even consume their own waste over time. There are also a lot of questions about how waste streams from existing nuclear technologies could be better dealt with.

What about safety? My impression is that nuclear can actually be very safe, Chernobyl notwithstanding. Moreover, advancing technologies intersect in all sorts of ways that could make nuclear more feasible in *more places*, e.g., improved Earthquake prediction might help.

The basic idea of: make a surplus of cheap, safe, robust nuclear energy –> electrify many other sectors to reduce their GHG emissions –> also power negative emissions industrial direct air capture… seems like something we should be seriously considering.

But in practice, the high up front construction cost — currently perhaps $4000-$5000/kW, roughly 5x the construction cost of the other types of power plant shown above — could be a

concern, as could be the need to move beyond traditional Uranium light water reactors to get to full grid scale. Analyzing this is complex, as prices are changing and will change further with the advent of small modular reactors, and there will naturally be an ever-changing mixture of energy sources used and needed as renewables scale up to a significant fraction of the power grid.

Interlude: Fusion

Fusion is an interesting one. There is no fundamental obstacle, and many advantages over fission-based nuclear. Around mid-century, it will probably be working and becoming widespread. The key question is whether this can happen sooner, say in the next 15 years. Advantages, paraphrased and quoted from the ITER website, include:

- Can have roughly 4x energy per fuel mass over fission
- Abundant fuel, e.g., hydrogen isotopes deuterium and tritium. Deuterium can be distilled from water, then tritium generated ("bred") during operation, when neutrons hit lithium in the walls of the reactor: "While a 1000 MW coal-fired power plant requires 2.7 million tonnes of coal per year, a fusion plant of the kind envisioned for the second half of this century will only require 250 kilos of fuel per year, half of it deuterium, half of it tritium."
- Less of a radioactive waste problem. Basically makes helium from isotopes of hydrogen. Particles hitting materials in the reactor do activate the materials in the reactor, but not creating large amounts of long-lived waste like in fission.
- Safety: "Only a few grams of fuel are present in the plasma at any given moment. This makes a fusion reactor incredibly economical in its fuel consumption and also confers important safety benefits to the installation." Also, "A Fukushima-type nuclear accident is not possible in a tokamak fusion device. It is difficult enough to reach and maintain the precise conditions necessary for fusion—if any disturbance occurs, the plasma cools within seconds and the reaction stops. The quantity of fuel present in the vessel at any one time is enough for a few seconds only and there is no risk of a chain reaction."

Otherwise, pretty similar to fission in terms of what it would look like for scale-up to power the world. These are big complex machines. It is a very difficult way to boil water, and not clear how well cost will scale down, unlike say, solar cells.

No demonstration reactor has yet generated net energy gain. ITER is expected to get 10x gain for a duration of several minutes. It will also be useful for testing all sorts of control mechanisms and properties of the plasma and its effects. But this will be happening after 2035 according to their plan, from what I understand.

The main approach now being developed uses magnetic confinement of a plasma that must be heated to >100 million degrees Celsius. Tokamaks are one class of magnetic confinement system, in which the plasma follows a helical magnetic field path inside a toroidal shaped vacuum chamber.

ITER is a Tokamak. ITER is a big, long, complex international project that is deliberately distributing effort across many international organizations. It extends well past 2040 for its major scientific milestones, not to mention future power plant scale followups and commercialization. In contrast, a previous fusion reactor, JET was completed much faster, decades ago.

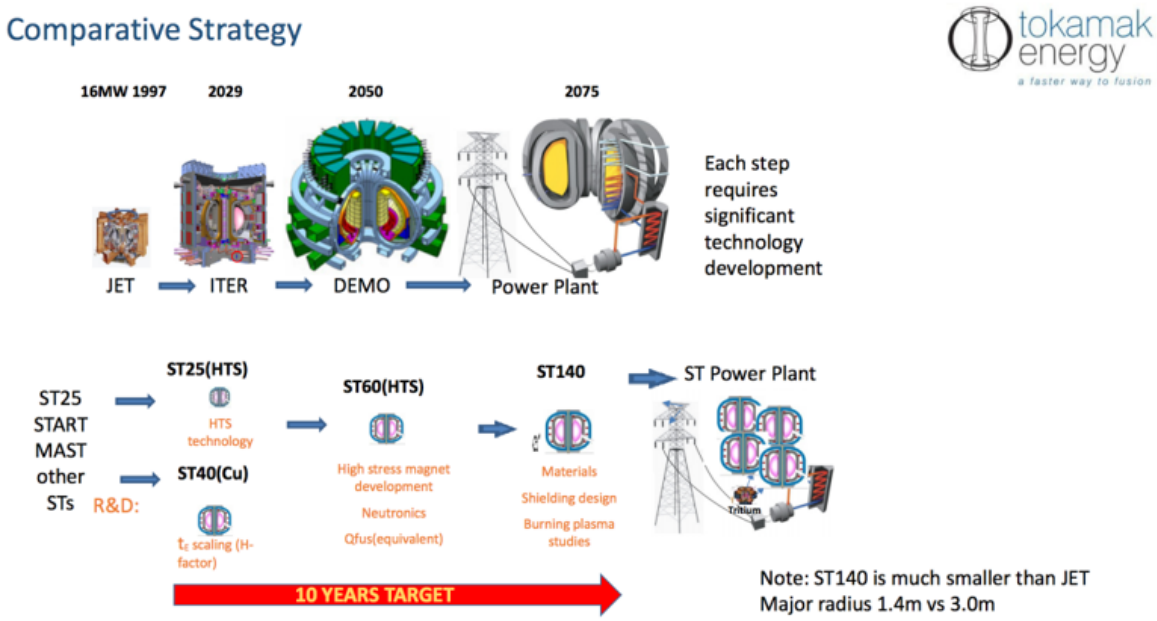
A recent National Academies Report explained why, though it is important to stay involved in ITER, new opportunities could allow a faster and somewhat cheaper approach to a full integrated pilot plant, which would hopefully synergize well with ITER. What makes sense in addition to ITER is pushing "compact fusion". But you have to understand that this is not easy or fast, either. You need things like: creating walls that can withstand a huge bombardment of neutrons, and manufacturing large amounts of high-temperature superconducting wires or tapes to make compact yet very-high-field magnets.

From the NAS report, the challenges include the following: “A compact fusion pilot plant requires developing operation scenarios for sustaining high-power density burning plasma with the plasma exhaust capability required for compact fusion. The engineering *design and fabrication of large bore, high-field superconducting magnets* for fusion needs to be established. *Long-lifetime materials* will need to be developed and qualified for use in the compact fusion pilot plant. *Tritium science and fusion breeding blanket development* needs to be developed sufficiently for integrated nonnuclear testing of prototypes that can serve as the basis for blanket components that will ultimately be installed in the compact pilot plant. Finally, to be successful, a detailed *system engineering* effort is needed to guide a “pre-pilot-plant” research program toward construction of a low capital-cost fusion pilot plant”.

Note the importance of large scale manufacturing of high temperature superconducting wires and tape: “Today’s opportunity for compact magnet fusion energy results from the potential for high-field superconducting magnets. New *high-temperature superconductors may make possible fusion magnets that can achieve fusion gain and power equivalent to ITER but at a significantly lower size and cost*… Industry can now produce *commercial quantities of high-temperature superconducting tapes* that have the potential to be used in very high field fusion magnets that will reduce the size needed to magnetically confine a burning plasma.”

So it will take at minimum years, if not a decade or more, perhaps much longer, just to get to a compact fusion pilot plant, which may not even be strongly net-positive in energy gain, let alone full scale useful power plants with economies of scale and proven industrial rollout: “The initial pilot plant operation would demonstrate net-electric equivalent performance in a compact fusion system, focusing on integrated core/edge performance, assessing plasma material interactions, demonstrating tritium pumping, limited breeding, safe handling, and extraction. This initial phase *would not include long-pulse fusion power production and would not demonstrate self-sufficient tritium production*. The second phase of the pilot plant would seek near continuous operation, allow for materials/component testing with neutron fluences approaching power-plant levels, and provide integrated blanket testing. Upon success of this second phase, *the compact fusion pilot plant studies would have reduced both the economic and technical risks for fusion energy-based electricity production* and will motivate further involvement from industries and utilities…”

There are a number of companies pursuing things that are along these lines of some kind of next-gen compact magnetic confinement, including Commonwealth Fusion, Tokamak Energy, and Lockheed Martin’s Skunkworks. Here is a figure from some slides online from Tokamak Energy describing their ambitious timeline (“ST” indicates “spherical tokamak” and HTS means “high temperature superconductor”):



https://stfc.ukri.org/files/uk-magnetics-tokamak-energy/

There is also a need to move beyond this framework to include other novel ways of confining the plasma. A distinct hybrid approach is being pursued by TAE. There was an ARPA-E program called ALPHA to pursue initial academic-scale research on hybrid approaches that move beyond classical magnetic confinement type approaches, including so-called magneto-inertial fusion and "z-pinch". Here is from the ALPHA website: "Most mainstream fusion research currently focuses on one of two approaches to confining plasmas: magnetic confinement, which uses magnetic fields and lower-than-air ion densities, and inertial confinement, which uses heating and compression and involves greater-than-solid densities. The ALPHA program aims to create additional options for fusion research by developing the tools for new, lower-cost pathways to fusion, and with a focus on intermediate densities in between these two approaches. These new intermediate density options may offer reduced size, energy, and power-density requirements for fusion reactors and enable low-cost, transformative routes to economical fusion power." The company General Fusion appears to be pursuing something along those lines, and the company Helion was in the ALPHA portfolio according to the just-linked retrospective. Todd Rider, circa 1990s, has lots of broader ideas and critiques of the fusion field, which may have pointed to a need for more novel designs.

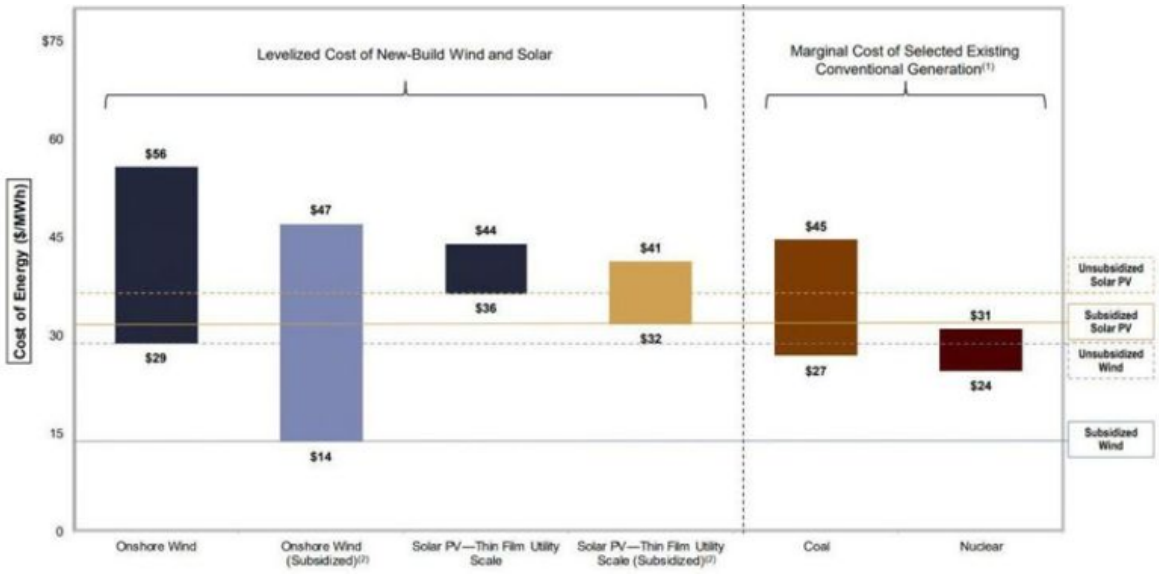
It seems like US fusion funding stalled in the 1990s and onward, and is only starting to be rebooted. I have no idea of whether it will indeed be possible to demonstrate a full net-positive energy gain pilot plant, let alone really commercialize, in the next 10-15 years, but it may happen if pushed very aggressively and with some luck.

Variable renewables

There is a project at Stanford to outline paths to fully renewable energy by 2050 for 139 countries around the world, without nuclear.

Subsidies have helped drive down the cost of wind and solar energy to the point where they are cost-competitive with fossil fuels (in some settings and up to some level of scale) — this is fantastic, as it means it is feasible to massively build them out now, which we should do. As I mentioned in the disclaimers, deploying existing technology now is important as there is a tightly limited feasible carbon budget in the coming few decades and we can't afford to wait to curb emissions if we want to limit the peak CO2 concentration reached. Arguably, the scenarios in integrated assessment models may not be sufficiently accounting for how fast

prices can drop as technologies like solar get deployed at larger scales. Here is a nice article from Megan Mahajan on the plunging price of renewables, building on studies by Lazard, and there are other such studies here: the results from Lazard are shown in the following figure



forbes.com/sites/energyinnovation/2018/12/03/plunging-prices-mean-building-new-renewable-energy-is-cheaper-than-running-existing-coal/

forbes.com/sites/energyinnovation/2018/12/03/plunging-prices-mean-building-new-renewable-energy-is-cheaper-than-running-existing-coal/

Overall, though, variable renewables like wind and solar are still a small fraction of overall power generation. While subsidies for renewables have allowed scaling up of deployment and thus driven down their un-subsidized costs, there remain real obstacles to scaling them *to the level of the entire power grid*: even if we had much more low-cost renewables available, their intermittent and variable nature makes them difficult to integrate into the existing power grid, beyond a certain fractional level, without some form of stabilization. This could include large-scale storage, improved transmission systems that allow balancing the grid through integration of more geographically diverse and hence uncorrelated sources (e.g., via the technology of "high voltage direct current" transmission), flexible long-distance digital control of certain loads, as well as novel distributed control strategies.

In a system nearing 100% variable renewable energy without storage, the control architecture of the power grid has to change, because existing control strategies rely heavily on the mechanical inertia of the large spinning turbines in conventional power plants — the current small percentage of variable renewables in the grid is able to integrate by following the patterns established by these large non-variable sources, and by being smoothed out by conventional fossil fuel sources. The Stanford roadmap makes reference to a separate study where they did detailed simulations of variability based on weather, and incorporated the use of various existing storage technologies to allow grid integration. A paper was published in PNAS, however, questioning their assumptions and arguing:

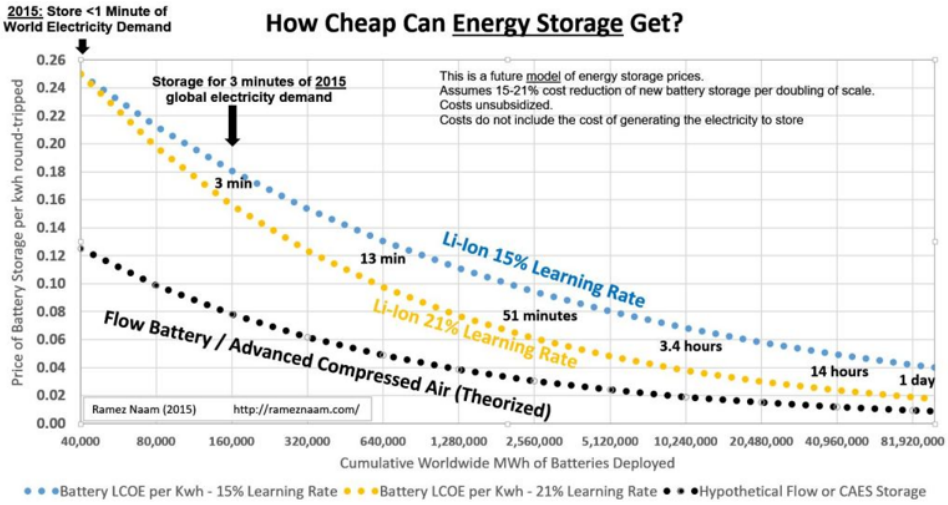
> *"... it is not in question that it would be theoretically possible to build a reliable energy system excluding all bioenergy, nuclear energy, and fossil fuel sources. Given unlimited resources to build variable energy production facilities, while expanding the transmission grid and accompanying energy storage capacity enormously, one would eventually be able to meet any conceivable load. However, in developing a strategy to effectively mitigate global energy-related CO2 emissions, it is critical that the scope of the challenge to achieve this in the real world is accurately defined and clearly*

communicated... The study... assumes a total of 2,604 GW of storage charging capacity, more than double the entire current capacity of all power plants in the United States... The ***energy storage capacity consists almost entirely of two technologies that remain unproven*** *at any scale: 514.6 TWh of UTES (the largest UTES facility today is 0.0041 TWh) (additional discussion is in SI Appendix, section S2.1) and 13.26 TWh of phase change materials (PCMs; effectively in research and demonstration phase) (additional discussion is in SI Appendix, section S2.2) coupled to concentrating solar thermal power (CSP). To give an idea of scale, the 100% wind, solar, and hydroelectric power system proposed in ref. 11 envisions UTES systems deployed in nearly every community for nearly every home, business, office building, hospital, school, and factory in the United States, although only a handful exist today..."*

The above-linked post from Uhlik on thorium reactors also contains a quantitative critique of the assumptions of the Jacobson wind-water-solar roadmap.

(I would be remiss not to mention the Stanford group's reply which among other things states: "Clack et al. (1) assert that underground thermal energy storage (UTES) can't be expanded nationally, but we disagree. UTES is a form of district heating, which is already used worldwide (e.g., 60% of Denmark); UTES is technologically mature and inexpensive; moreover, hot-water storage or heat pumps can substitute for UTES. Similarly, molten salt can substitute for phase change materials in CSP storage.")

How much storage does one really need to go fully to variable renewables, and what would the cost of storage have to be for this to be possible? This paper says the cost of storage would have to go to about $20/kWh for a cost-competitive 100% variable renewables grid. Meanwhile, lithium ion batteries currently cost around $140/kWh, but are projected to reach $60/kWh by 2030: see projection from Ramez Naam below. However, the paper just mentioned *also* pointed out that, if you only need to be 95% based on variable renewables, then the storage cost target is significantly less stringent. That seems fairly optimistic for the medium term.



From Ramez Naam: http://rameznaam.com/2015/10/14/how-cheap-can-energy-storage-get/

From Ramez Naam: http://rameznaam.com/2015/10/14/how-cheap-can-energy-storage-get/

Can we ballpark why these kinds of optimistic numbers could make some intuitive sense? Ramez Naam gives some explanation here and here. Here's another quick stab at it, building on Naam and MacKay: Suppose your lithium ion battery costs $100 per kWh of capacity (which will be true in a few years) and can be charged/discharged 1000 times. That's 10 cents to charge/discharge a kWh of battery each time — you can see that's the range in the middle of Ramez Naam's curve above. Now, that's comparable to the raw electricity cost, so

if you needed to be constantly charging and discharging the battery to supply every kWh of electricity that you use, you'd be doubling your energy cost. But ideally you only need to use the battery infrequently. Let's say that a person in a rich country needs to be able to store 20 kWh (a Tesla car battery is ~60-80 kWh) per person over periods on the order of 5 days (this last number is from MacKay for the case of wind power in England). Your 20 kWh charge/discharge costs $2, and let's say you have to do this once a week. That's $104 dollars a year on your energy bill due to the battery, likely only a fraction of your total electricity cost if you're in a rich country.

Still, that's a lot of batteries. Eventually, it seems limited world lithium reserves will become a problem. But of course, there may be other ways to make batteries based on other materials, and new forms of storage, that can kick in by then.

(MacKay argues that the best variable renewables grid integration approach for England would indeed be to massively deploy electric cars, and to make their charging process smart so that they would charge primarily when there is a surplus of power available and not charge when there was a deficit. You'd need to get a large fraction of people having a smart-charging electric car for that to work: he estimates a fleet of 30 million such grid-synced smart-charging electric cars could allow England to be powered by wind. That's about half of all people in England owning an electric car. Even better would be if they would donate power back to the grid as opposed to simply adjusting demand to better match supply. Smart buildings and fridges are also in this category.)

A brief interlude on other grid-scale storage technologies:

- *Compressed air energy storage* has some very appealing features, e.g., it can be very large-scale, and doesn't require any exotic materials like lithium. Here is the basic issue. When you compress air quickly, you heat it. This heat has to go somewhere and then has to be recovered. Otherwise you have thermal losses. Ideally, you'd be able to do isothermal compression and expansion, shuttling heat slowly to and from the environment at a constant temperature with no energy permanently lost in doing so — any heat you give off, you get back for free. Unfortunately this is not easy to do for compressed air energy storage, e.g., you have to compress a lot of air to a high pressure quickly, and this is hard to do while maintaining a constant temperature and fast thermal contact with the outside world. Innovative but now sadly defunct company LightSail tried to deal with this through a "quasi-isothermal" process in which water vapor was injected into the compressing air to absorb the heat and then release it back during expansion — because water has a very high specific heat, it can absorb a lot of heat with minimal change in temperature. Fortunately, there are published papers and patents and others taking related approaches. How do the numbers work out for this? Using the formula for the energy associated with isothermal compression, we could store in theory about 50 kWh per cubic meter of compressed air at 300 atm pressure, whereas LightSail was trying to reach about 30 kWh per cubic meter. (Beyond the quasi-isothermal approach, there are others that try to be adiabatic or that couple the heat to things like hydrogen production.)
- *Molten salt* thermal energy storage is basically storing heat. You want to store a lot of heat, and for purely practical reasons melted salts are a good medium in which to store it. As this article explains: "The salt is a combination of sodium and potassium nitrate, with a melting temperature of 460°F. In the liquid state, molten salt has the viscosity and the appearance similar to water. "In solar applications, molten salt is used for a number of practical reasons," says Terry Murphy, Chief Executive Officer for SolarReserve, who along with others helped develop the molten salt technology at Rocketdyne. "Molten salt is a heat storage medium that retains thermal energy very effectively over time and operates at temperatures greater than 1000°F, which matches well with the most efficient steam turbines. Second, it remains in a liquid state throughout the plant's operating regime, which will improve long-term reliability and reduce O&M costs. And third, it's totally 'green,' molten salt is a non-toxic, readily available material, similar to commercial fertilizers." A primary advantage of molten

salt central receiver technology is that the molten salt can be heated to 1050°F, which allows high energy steam to be generated at utility-standard temperatures (1650 psi minimum, 1025°F), achieving high thermodynamic cycle efficiencies of approximately 40 percent in modern steam turbine systems…" Laughlin has proposed a particular heat engine design using such molten salts.

- *Flow batteries* are electrochemical energy storage. The "flow" refers to the fact that you don't have to place the chemicals right in the path between the electrodes: the chemicals can be in separate tanks that can flow into that space as needed. Also, they can be based on materials more abundant than lithium, e.g., iron, and even largely on organic chemicals, not rare elements like lithium. As this article explains: "Flow batteries store energy in chemical fluids contained in external tanks—as with fuel cells—instead of within the battery container itself. The two main components—the electrochemical conversion hardware through which the fluids are flowed (which sets the peak power capacity), and the chemical storage tanks (which set the energy capacity)—may be independently sized. Thus the amount of energy that can be stored is limited only by the size of the tanks. The design permits larger amounts of energy to be stored at lower cost than with traditional batteries."

While it seems possible to achieve a nearly 100% variable renewables grid with massive deployment of conventional batteries and/or lots and lots of electric cars, achieving breakthroughs in other forms of ultra-low-cost renewables storage — with longer working lifetimes, simpler manufacturing, less stringent materials requirements, higher capacities and lower costs than batteries when used at grid scale (e.g., Laughlin's paper on Malta suggests a marginal cost of $13/kWh) — would make the grid integration and stability problems simpler and cheaper to deal with.

It is therefore not yet clear to me exactly how much hinges on simply bringing down the raw cost of solar or wind generation through deployment and hence economies of scale of existing technologies, versus scaling up the manufacturing and use of existing storage technologies such as lithium ion and similar batteries, versus novel R&D on storage and grid integration to solve the problems that arise once the variable renewables penetration goes above 50% or so, versus pushing existing or new nuclear, geothermal, and CCS technologies. Probably a mixture makes it all more tractable.

In addition to storage per se, surplus energy generated when supply is high but demand is low may also be used for applications like desalination of water or hydrogen fuel production via electrolysis; hydrogen in turn has a number of applications to decarbonizing other sectors.

Here's something cool that connects all these issues: a company making a nuclear reactor that runs on existing stockpiles of spent nuclear fuel and that is expressly designed to provide *tunable* output to be used to buffer variability in *renewables* and thus speed the integration of renewables into the grid — nuclear and variable renewables not competing but collaborating.

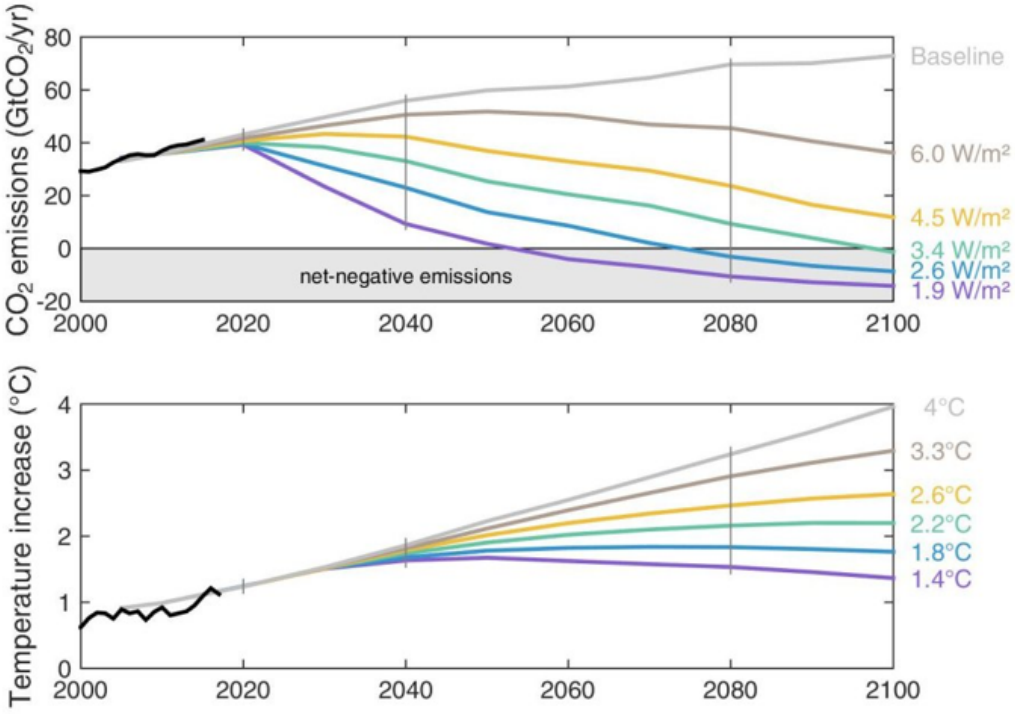
How quickly must we fully decarbonize?

In future posts, I'll summarize some of the literature on other aspects of possible future intervention, including 1) *carbon dioxide removal*, i.e., *negative emissions* technologies — needed in current scenarios for staying under 1.5C and becoming less and less controversial — as well as 2) much more controversial potential ways to disrupt the coupling between CO2 concentration and temperature (solar geo-engineering) that would likely only be used on an emergency basis and would pose difficult governance challenges.

But quickly decarbonizing the electricity sector and, crucially, the other large sectors —steel and cement production, agriculture, and land use / deforestation / desertification among others — is still the primary need.

How quickly? The details are complex, vigorously debated and beyond what I can hope to effectively cover. It looks like we want to get to *net-negative* CO2 emissions by around 2050-2080 to have some appreciable likelihood of staying within 2C peak warming, and by 2050 proper for 1.5C.

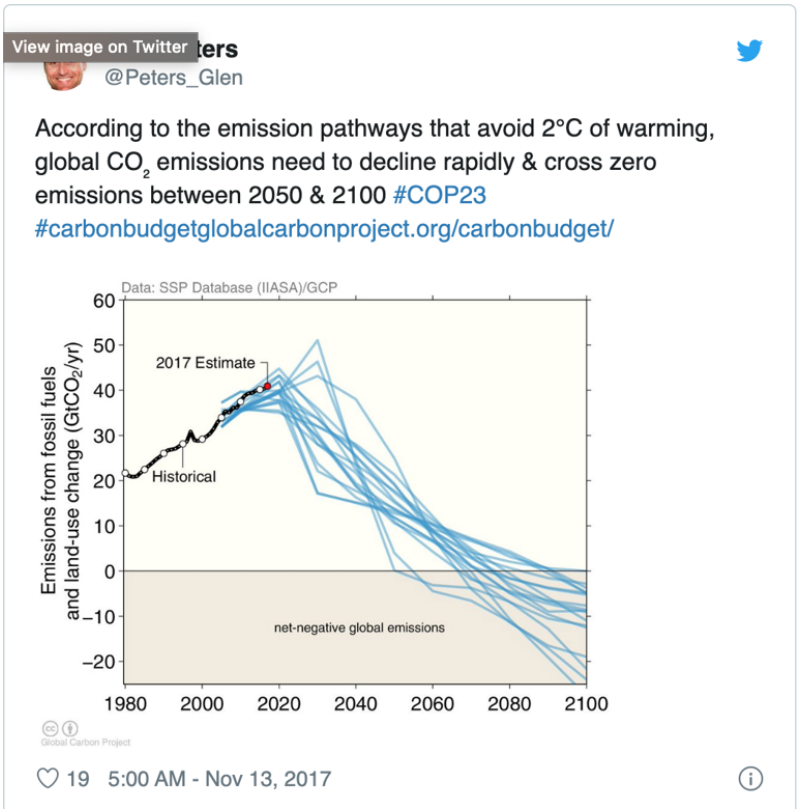
Here is one example graph from Glen Peters on Twitter, with emissions scenarios (top) and their associated global average surface temperatures (bottom) — and be careful to look at the peak temperature along the lower graph, not just the value reached in 2100:



From Glen Peters: https://mobile.twitter.com/peters_glen/status/1164441790386188288

From Glen Peters: https://mobile.twitter.com/peters_glen/status/1164441790386188288

And here is another (allowing a potential temperature overshoot mid-century I believe):



https://twitter.com/Peters_Glen/status/930042574777257984

(Arguably, talking about “net zero” is too vague and we should instead have separate, definite targets for both positive and negative emissions.)

The need for a (revenue-neutral, staged) carbon tax

How can this change be incentivized? There are powerful arguments in favor of a carbon tax. Specifically, one can do a gradually scaled-up, revenue-neutral carbon tax that would, for instance, return the revenues back to consumers as rebates or tax reductions in other kinds of taxes — this could be important as almost any form of carbon tax would ultimately be passed at least in significant part onto the consumer, and many consumers are cash-strapped already, e.g., struggling to afford gas for their commute.

How big a carbon tax, and what effects would that have on citizens? That’s a complicated question and not one I have found a full answer to. Typical values suggest something like $10-$100 per tonne of CO2, meant to offset widely-varying estimates of the social cost of carbon. That would apparently increase, say, the gas bill for your car by roughly 10 cents to 1 dollar per gallon of gas (the current average price of gas in the USA is around $2.50/gallon, and there is already an excise tax on it of about 1/10 that).

Alas, consumers in the USA seem wary of changes in that range. It is unclear if they would *still* be wary if they were presented with a rebate (remember, the carbon tax can be revenue-neutral, or at least with significant offsets to give cash back to citizens, although some would hope to use revenues from a carbon tax to fund climate-related programs and deployments) that would o*n net leave most people’s bank accounts the same or fatter*, not thinner.

A few points of comparison:

- We'll see in the next post that technologies aiming to scrub carbon from the atmosphere are shooting to get their cost of CO2 produced into this same range, say <$100/tonne CO2.
- If the US is currently emitting 6 GigaTonnes of CO2 and we had a $40/tonne CO2 carbon tax right now, that would presumably be something like $240 billion per year in carbon tax revenue, or a bit under 1/10 of all US federal tax revenue of $3.something trillion. Could we re-structure the rest of the tax code to give back 1/10 to consumers, as a rebate, so that their bank accounts (and the overall size of government) stay roughly flat? I bet we could. But this puts into perspective that, if someone is instead talking about a $400/tonne CO2 price of carbon (see the Stern Review item in the table below), that's approaching the totality of all other tax revenue in the USA at current emissions levels.
- That $240 billion paid against the $40/tonne carbon price is around 1.2% of US GDP of around $20 trillion.

There are also some arguments against a carbon tax, e.g., that a carbon tax in the US doesn't impact other countries like China and may simply shift international trade balances, or that they might be revoked, as well as the need for them to be sufficiently large and thus correspondingly painful to the consumer in order to make a difference. They are also empirically not easy to pass, and economists are studying why.

France had a fuel tax of about $60/ton CO2 equivalent, but when Macron tried to double that over the next 4 years, the Yellow Vest protests happened (this may of course be specific to the kind of fuel tax they imposed and the particular situation there). Canada did pass a carbon tax starting at below that level and rising to near it over a few years ("start low at $20 per ton [tonne of emitted CO2] in 2019, rising at $10 per ton per year until reaching $50 per ton in 2022"). Recently, some major fossil energy companies started lobbying in favor of a carbon tax, which I found surprising — alas, it looks like they are putting in pennies compared to their fossil-fuel-promoting endeavors, in a bid to insure themselves against future climate-related lawsuits!

Cap and trade is another possibility, arguably fundamentally similar. Cap and trade has the advantage of *directly* setting a specific emissions limit. China has established something of a cap and trade system. It doesn't look that stringent, yet: "The scheme set the initial carbon allowances to 3–5 billion tonnes per year."

Subsidizing R&D into decarbonization technologies is the key alternative suggested by Noah Smith — if you get a breakthrough towards lower-cost technology, that can be adopted anywhere in the world (including places that may not feel they can afford a carbon tax), and only benefits people (no change in tax structure). We obviously need that too, big time. This thread has a coherent discussion about why carbon taxes can help to scale existing technologies but, for areas where truly new technologies are needed (of which there are many), targeted R&D and highly specific subsidies can be better.

Before we finish this first post, let's take a look at how economists estimate how big a carbon tax we really should have? Consider this snippet from the Nobel Prize announcement for economist William Nordhaus, who pioneered this area:

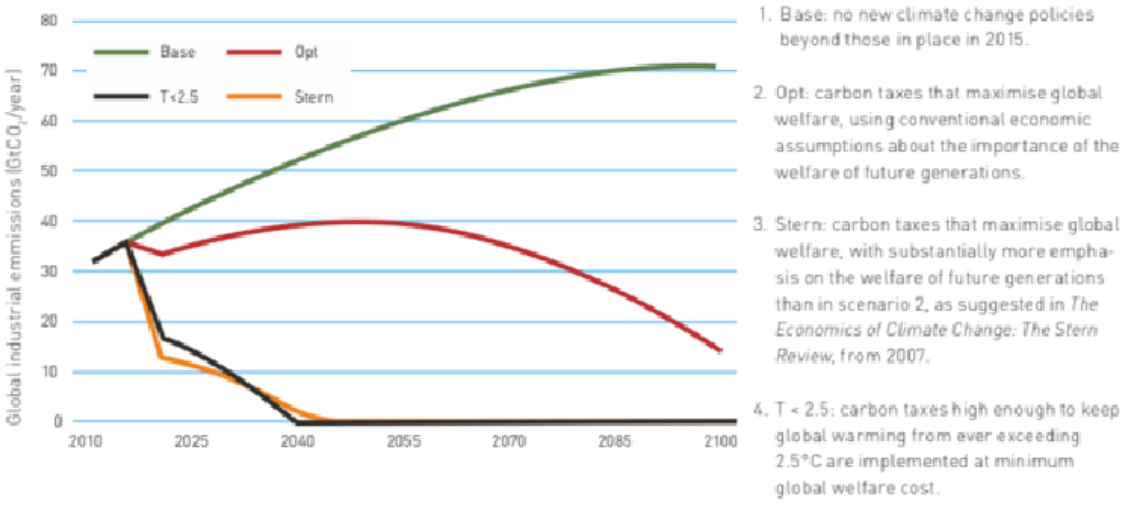


Figure 3: CO_2 emissions over time for four climate policies (explanations in the text). Predictions from the DICE-2016R2 model, according to Nordhaus' own simulations.

Figure 3 shows CO_2 emissions over time in each of these four scenarios. The different paths for carbon taxes mean that emissions – and thus the extent of climate change – are very different across these scenarios. In scenario 2, taxes start out at around 30 USD/ton CO_2 and rise over time at about the same rate as global GDP. In scenarios 3 and 4, which have much more drastic emission cuts, taxes are 6-8 times higher.

https://www.nobelprize.org/uploads/2018/10/popular-economicsciencesprize2018.pdf

https://www.nobelprize.org/uploads/2018/10/popular-economicsciencesprize2018.pdf

Nordhaus uses coupled climate + economic activity and growth models to evaluate "optimal" scenarios — ones with the maximum overall growth in the economy, factoring in the effects of damage from climate change and the costs of introducing decarbonization approaches. His *optimal* overall economic scenario ("Opt") is the one shown in red in the plot above, with a global carbon tax starting at $30-40/tonne CO2 or so (and rising to about $50/tonne by 2030 and then further afterwards, see table below). Note, though, that this pathway massively overshoots the orange and black pathways in terms of emissions — the black and orange pathways look more like the other graphs we've seen for *full decarbonization by 2050*. Here is a table from his 2017 PNAS paper with a comparison:

Table 1. Global SCC by different assumptions

Scenario	Assumption	2015	2020	2025	2030	2050
Base parameters						
	Baseline*	31.2	37.3	44.0	51.6	102.5
	Optimal controls†	30.7	36.7	43.5	51.2	103.6
2.5 degree maximum						
	Maximum†	184.4	229.1	284.1	351.0	1,006.2
	Max for 100 y†	106.7	133.1	165.1	203.7	543.3
The Stern Review discounting						
	Uncalibrated†	197.4	266.5	324.6	376.2	629.2
Alternative discount rates*						
	2.5%	128.5	140.0	152.0	164.6	235.7
	3%	79.1	87.3	95.9	104.9	156.6
	4%	36.3	40.9	45.8	51.1	81.7
	5%	19.7	22.6	25.7	29.1	49.2

The SCC is measured in 2010 international US dollars.
*Calculation along the reference path with current policy.
†Calculation along the optimized emissions path.

Nordhaus, William D. “Revisiting the social cost of carbon.” Proceedings of the National Academy of Sciences 114.7 (2017): 1518-1523.

Nordhaus, William D. “Revisiting the social cost of carbon.” Proceedings of the National Academy of Sciences 114.7 (2017): 1518-1523.

Indeed, the optimality of an Opt-like pathway is a subject of great controversy. Nordhaus’s “optimal” carbon pricing, in that model, would get us to a temperature rise of 3.5C (!) or so from preindustrial, while pushing to full decarbonization by 2050, in order to stay below 2C, would require roughly an order of magnitude higher carbon price. Staying below 1.5C would seem, according to such models, to imply a need for even higher carbon prices. From a recent review:

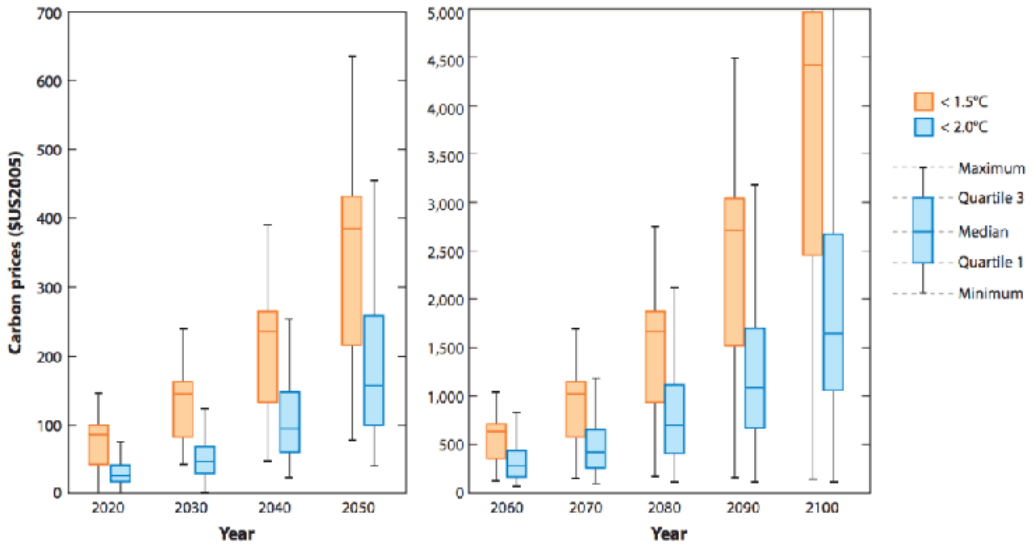


Figure 2

Carbon prices ($2005) in 1.5°C and 2°C scenarios. Carbon prices for the <1.5°C scenarios were computed from a scenario set obtained by pooling the 37 scenarios in Reference 99 and the five scenarios in Reference 187 that have a >50% probability of limiting warming to 1.5°C by 2100. Carbon prices for the <2°C scenarios were computed from the 125 scenarios in Reference 187 with a >50% probability of limiting warming to between 1.75°C and 2°C.

Dietz S, Bowen A, Doda B, Gambhir A, Warren R. The economics of 1.5 C climate change. Annual Review of Environment and Resources. 2018 Oct 17;43:455-80.

Dietz S, Bowen A, Doda B, Gambhir A, Warren R. The economics of 1.5 C climate change. Annual Review of Environment and Resources. 2018 Oct 17;43:455-80.

Now, obviously, if Nordhaus’s model had included, say, Antarctica completely melting and many major cities going completely underwater due to a series of major sea ice and other tipping points being crossed, the social cost of carbon would need to be a lot higher. So implicitly the risk of such scenarios is considered low in such models. At a more subtle level, the particulars of these optima depend on the expected cost of decarbonization (which depends on the rate of technological innovation), on the temporal discount rates used (you can see in the table above that prices rapidly rise with decreasing discount rate), and on other factors.

There are some good-sounding arguments for having carbon prices instead start high, and then decline, political difficulties notwithstanding — I haven’t understood the full details, but this approach seems to focus on mitigating risk and to take into account more of the variance in scenarios and our finite but improving knowledge thereof.

The fact is, as the Nobel Prize announcement notes, there are huge uncertainties in what would really be optimal (even *within* the philosophical frame that underlies Nordhaus's analysis, e.g., with the type and rough magnitude of the discount rate he uses, and even neglecting the politics of passing such taxes) — it depends on the cost of decarbonizing quickly, and on the climate damage cost of *not* decarbonizing quickly, both of which are highly uncertain at present. So we have to take a best guess — but that best guess should incorporate the uncertainty.

This brings us back to the role of technology: it can *reduce the cost of decarbonization*, and it can *mitigate tail risks* from climate damage. Advanced technology is crucial to allow us to decarbonize safely and quickly, with minimal economic cost. It allows a lower self-imposed carbon price to have greater benefit, shifting the estimated-to-be-economically-optimal pathways to overlap those that strongly limit the risk of crossing catastrophic climate tipping points.

Tentative conclusions

My sense — again, as an outsider to the field simply trying to grok the literature in my free time — is that we should do the following:

- Make sure our ability to measure and model everything in the climate system continues to improve, including for complex land and ocean biological effects.
- Lobby for a revenue-neutral, staged price on carbon, in rich countries initially — rich countries can bring down the cost for poor countries by developing and scaling up better, cheaper technology. In the USA there is a specific program with some high-profile advocates called the Carbon Dividends Plan. I certainly don't want to say that a carbon tax is the only necessary policy though — also advanced R&D, tax credits, subsidies, lobbying reform, and down the line hopefully new social technology for making societal decisions would come into play.
- Economically incentivize a full transition to net-negative emissions by around 2050. This includes deploying a portfolio of existing clean technologies now to limit our cumulative carbon emissions and peak CO2 concentration.
- Heavily invest in large-scale, advanced R&D to bring down the cost of grid-scale energy storage, next-generation nuclear, novel clean manufacturing processes for things like steel and cement, improved agricultural technologies, smart grid-integrated buildings, and synthetic meat (among other areas). With enough technological innovation, I hope that the cost to decarbonize could be brought lower than the estimates of those like Nordhaus.
- Heavily invest in sustainable land use management and conservation, as well as aggressive afforestation.
- Switch to electric cars. (This will impact multiple key areas.)
- Bootstrap the nascent fields of direct air carbon capture / carbon sequestration, both chemical and biological, to the point where they can really scale cost-effectively (next post), and seriously consider ocean liming, sequestration of agricultural waste biomass, and so forth (also next post). This involves both new research and the creation of new markets for carbon. Use sequestered carbon to make the jet fuel.
- Do the fundamental R&D on emergency adaptation strategies like solar radiation management and buttressing of glaciers (next next post).

You can go to the next post on negative emissions.

A simple sketch of how realism became unpopular

[Epistemic status: *Sharing current impressions in a quick, simplified way in case others have details to add or have a more illuminating account. Medium-confidence that this is one of the most important parts of the story.*]

Here's my current sense of how we ended up in this weird world where:

- I still intermittently run into people who claim that there's no such thing as reality or truth;
- a lot of 20th-century psychologists made a habit of saying things like 'minds don't exist, only behaviors';
- a lot of 20th-century physicists made a habit of saying things like 'quarks don't exist, only minds';
- there's a big academic split between continental thinkers saying (or being rounded off to saying) some variant of "everything is culture / perception / discourse / power" and Anglophone thinkers saying (or being rounded off to saying) "no".

Background context:

1. The ancient Greeks wrote down a whole lot of arguments. In many cases, we're missing enough textual fragments or context that we don't really know why they were arguing — what exact propositions were in dispute, or what the stakes were.

2. In any case, most of this is screened off by the fact that Europe's memetic winners were Christianity plus normal unphilosophical beliefs like "the sky is, in fact, blue".

3. Then, in 1521, the Protestant Reformation began.

4. In 1562, the Catholics found a giant list of arguments *against everything* by the minor Greek skeptic Sextus Empiricus, got very excited, and immediately weaponized them to show that the Protestant arguments fail (because all arguments fail).

5. These soon spread and became a sensation, and not just for being a useful superweapon. Plenty of intellectuals were earnest humanists used to taking arguments at face value, and found Sextus' arguments genuinely upsetting and fascinating.

I trace continental thinkers' "everything is subjective/relative" arguments back to a single 1710 error in George Berkeley:

> [...] I am content to put the whole upon this Issue; if you can but conceive it possible for one extended moveable Substance, or in general, for any one Idea or

> any thing like an Idea, to exist otherwise than in a Mind perceiving it, I shall readily give up the Cause[....]
>
> But say you, surely there is nothing easier than to imagine Trees, for instance, in a Park, or Books existing in a Closet, and no Body by to perceive them. I answer, you may so, there is no difficulty in it: But what is all this, I beseech you, more than framing in your Mind certain Ideas which you call Books and Trees, and the same time omitting to frame the Idea of any one that may perceive them? But do not you your self perceive or think of them all the while? This therefore is nothing to the purpose: It only shews you have the Power of imagining or forming Ideas in your Mind; but it doth not shew that you can conceive it possible, the Objects of your Thought may exist without the Mind: To make out this, it is necessary that you conceive them existing unconceived or unthought of, which is a manifest Repugnancy.

If I can imagine a tree that exists outside of any mind, then I can imagine a tree that is not being imagined. But "an imagined X that is not being imagined" is a contradiction. Therefore everything I can imagine or conceive of must be a mental object.

Berkeley ran with this argument to claim that there could be no unexperienced objects, therefore everything must exist in some mind — if nothing else, the mind of God.

The error here is mixing up what falls inside vs. outside of quotation marks. "I'm conceiving of a not-conceivable object" is a formal contradiction, but "I'm conceiving of the concept 'a not-conceivable object'" isn't, and human brains and natural language make it easy to mix up levels like those.

(I can immediately think of another major milestone in the history of European thought, Anselm's ontological argument for God, that shows the same brain bug.)

Berkeley's view was able to find fertile soil in an environment rife with non-naturalism, skeptical arguments, and competition between epistemic criteria and authorities. Via Kant and Kant's successors (Hegel chief among them), he successfully convinced the main current of 19th-century European philosophy to treat the idea of a "mind-independent world" as something ineffable or mysterious, and to treat experiences or perspectives as fundamental.

(**Edit:** G.E. Moore seems to think that everyone in the 19th century was making an error along these lines, but I now suspect Kant himself wasn't making this mistake; I think his main error was trying too hard to defeat skepticism.

I also don't think Berkeley's writing would have been sufficient to confuse Europe on its own; it's too lucid and well-articulated. The transition to Kantian and Hegelian versions of these arguments is important because they were much more elaborate and poorly expressed, requiring a lot of intellectual effort in order to spot the inconsistencies.)

My unscholarly surface impression of the turn of the 20th century is that these memes ("the territory is fundamentally mysterious" and "maps are sort of magical and cosmically important") allowed a lot of mysticism and weird metaphysics to creep into intellectual life, but that ideas like those are *actually* hard to justify in dry academic prose, such that the more memetically fit descendants of idealism in the 20th century ended up being quietist ("let's just run experiments and not talk about all this weird 'world' stuff") or instrumentalist / phenomenalist / skeptic / relativist ("you can't know

'world' stuff, so let's retreat to just discussing impressions; and maybe you can't even know those, so really what's left is power struggles").

Today, the pendulum has long since swung back again in most areas of intellectual life, perhaps because we've more solidly settled around our new central authority (science) and the threats to centralized epistemic authority (religious and philosophical controversy) are more distant memories. Metaphysics and weird arguments are fashionable again in analytic philosophy; behaviorism is long-dead in psychology; and quietism, non-realism, and non-naturalism at least no longer *dominate* the discussion in QM, though a lot of Copenhagen slogans remain popular.

The above is a very simple picture featuring uneven scholarship, and history tends to be messier than all that. (Ideas get independently rediscovered, movements go one step forward only to retreat two steps back, etc.) Also, I'm not claiming that everyone endorsed the master argument as stated, just that the master argument happened to shift intellectual fashions in this direction in a durable way.

What funding sources exist for technical AI safety research?

What funding sources exist? Who are they aimed at - people within academia, independent researchers, early-career researchers, established researchers? What sort of research are they aimed at - MIRI-style deconfusion, ML-style approaches, more theoretical, less theoretical? What quirks do they have, what specific things do they target?

For purposes of this question, I am not interested in funding sources aimed at strategy/infrastructure/coordination/etc, only direct technical safety research.

Does the US nuclear policy still target cities?

The history of nuclear strategic bombing

Daniel Ellsberg's *The Doomsday Machine* brought my attention to a horrifying fact about early US nuclear targeting policy. In 1961, the US had only one nuclear war plan, and it called for the destruction of every major Soviet city and military target. That is not surprising. However, the plan also called for the destruction of every major Chinese city and military target, *even if China had not provoked the United States*. In other words, the US nuclear war plan called for the destruction of the major population centers of the most populous country in the world, even in circumstances where that country had not attacked the United States or its allies. Ellsberg points out that at the time, people at RAND and presumably other parts of the US defense establishment understood that the Chinese and the Soviets were beginning to diverge in strategic interests and thus should not be treated as one bloc. Nevertheless, the top levels of the US command, including President Eisenhower, were committed to the utter destruction of both Chinese and Soviet targets in the event of a war with either country.

The policy of destroying cities is a legacy left over from strategic bombing in World War II. The destruction of Hiroshima and Nagasaki are the most famous, but the fire bombings of Japanese and Germany cities destroyed far more infrastructure and killed far more people than the two atomic bombs. The given rational for strategic bombing was to destroy the ability of the enemy states to continue to make war. If a state can no longer produce airplanes and tanks, either because the factories have been destroyed or because there are no longer people to work in the factories, then its ability to resist is diminished.

Given the level of technology and development in WWII, strategic bombing had a chance at achieving military objectives, because the conflict was to carry on for multiple years. On the timescale of years, a country's capacity to build armaments and resupply armies in the field can be crucial to victory.

Nuclear war changes this calculus. In a modern nuclear war involving SLBMs (Submarine launched ballistic missiles), ICBMs (Intercontinental ballistic missiles), strategic bombers, and other weapon systems, the majority of an adversary's military, industrial and population centers could be destroyed in a matter of days or hours. It is hard to imagine a nuclear war lasting years or even months. Without a prolonged war, the original rationale for strategic bombing disappears, or is at least much reduced. A state may still wish to reduce the capacity of its enemy to fight future wars, but it can no longer claim that the wholesale destruction of cities is necessary to achieve military objectives in the current war.

Why then, did early US nuclear policies call for the destruction of cities?

Nuclear game theory in the 1960s

The destruction of cities was primarily a threat of inflicting harm rather than an attempt to destroy the capacity of the enemy to wage war. The idea, formalized by RAND game theorist Thomas Schelling, was that both the United States and the Soviet Union would threaten massive retaliation against each other's civilian populations and industry to deter the other from starting a war.

Schelling developed a category of game theory that involved what he termed "mixed motive games". Games where both sides sought advantage, but where the payoff to one side did not strictly correlate to the loss to the other side. In these types of games, both players may wish to avoid outcomes that are mutually unfavorable (Strategy of Conflict, pg. 89). In the case of nuclear deterrence, both sides strongly preferred to avoid nuclear war, and thus were both were deterred from taking actions that would directly lead to nuclear war.

Much of Schelling's work concerns itself with how states in a nuclear stalemate can pursue their own advantage while avoiding escalation to nuclear war. In this type of game, states try to maneuver each other into positions where the only possible actions are 1) escalate and risk nuclear war or 2) de-escalate and concede something to the other side.

During the Cuban Missile Crisis, Kennedy ordered a blockade of Cuba, believing that such an action would not be sufficient for the Soviet Union to initiate a war. The United States believed that the Soviet Union would not try to break the blockade, because such an action would be recognized by both sides as starting a (nuclear) war. Because Kennedy proved correct in his belief that the Soviet Union would not go to war over the blockade nor risk initiating war by breaking the blockade, the United States used to its advantage both countries unwillingness to go to war.

What does this have to do with the targeting cities? To answer this question it's necessary to consider how a nuclear war might start. Although nuclear powers would almost always prefer to avoid a nuclear war, each has an incentive to strike first if they believe nuclear war to be inevitable. By striking first they may destroy their adversary's nuclear forces before they can be used. At this point, it's useful to define a couple of terms. Counterforce targeting refers to the targeting of enemy military installations, especially other nuclear forces. Countervalue targeting refers to the targeting of enemy infrastructure and population centers.

Consider the primary goal of a first strike. Under the most plausible nuclear war scenarios, it is to eliminate the nuclear forces of the rival state; its objective is primarily counterforce in nature. This is markedly different than the goal of a second strike. The primary goal of a second strike, under normal assumptions of deterrence, is actually to provide fulfilment of the pre-commitment made to retaliate if ever attacked. That is, it is necessary to actually be committed to attacking second so as to avoid being attacked in the first place.

Schelling effectively argued that the more punishing the second strike threatened to be, the more effective the deterrent would be also. If true, then in the event of a nuclear war, a state following the optimal strategy of deterrence would target cities as well as nuclear targets to make their nuclear response as punishing as possible; that is, it would destroy both counterforce and countervalue targets. This would seem to lead to a policy for states to attack cities, without a second thought. Indeed, this was the policy of the US and the Soviet Union in the 1950s and early 1960s. However, just because targeting cities promised to be a more effective deterrent did not mean it promised to be the best policy.

Given some probability of nuclear war, the effectiveness of a deterrent strategy ought to be weighed against the severity of the resulting war were that strategy employed. In other words, it might make sense for a state to commit to not targeting cities in a second strike if they themselves do not have their cities destroyed in a first strike. While this may reduce the effectiveness of their deterrent (and perhaps only marginally -- nuclear war is plenty damaging without cities being destroyed -- the fallout alone will kill many millions), it may also greatly reduce the severeness of a nuclear war.

Herman Kahn, a prominent and controversial RAND researcher, argued that states would be rational to refrain from destroying cities in a first strike, to retain some bartering power that might allow them to save more of their own cities. The argument is that the defending force might refrain from destroying many enemy cities if doing so prevented their own cities from being destroyed. Kahn believed that the US should study and prepare for negotiating for the avoidance of US cities in a nuclear war and that in order to do this the country should:

1. Develop the ability to have sufficiently protected or hidden nuclear forces to be able to both survive a first strike and carry out counterforce *and* countervalue attacks.
2. Have “backup presidents”, or people with authority to both order attacks and negotiate with the Soviet Union in the midst of a war, and that the US should have multiple secure locations which are staffed 24/7 by these leaders.

Both Herman Kahn and Thomas Schelling agreed that negotiating the end of a nuclear war would be difficult, but both believed it was critical that nuclear states remain capable of negotiation. Schelling writes about this in his 1966 work, Arms and Influence:

> *The closing stage, furthermore, might have to begin quickly, possibly before the first volley had reached its targets; and even the most confident victor would need to induce his enemy to avoid a final, futile orgy of hopeless revenge. In earlier times, one could plan the opening moves of war in detail and hope to improvise plans for its closure; for thermonuclear war, any preparations for closure would have to be made before the war starts.*
>
> *…A critical choice in the process of bringing a war to a successful close--or to the least disastrous close--is whether to destroy or to preserve the opposing government and its principal channels of command and communication. If we manage to destroy the opposing government’s control over its own armed forces, we may reduce their military effectiveness. At the same time, if we destroy the enemy government’s authority over its armed forces, we may preclude anyone’s ability to stop the war, to surrender, to negotiate an armistice, or to dismantle the enemy’s weapons.*

Historical developments in US nuclear targeting policy

The United State’s nuclear targeting policy has evolved from one of indiscriminate destruction of military and civilian targets, including cities, to one that promises proportional retaliation. While the public documents, perhaps intentionally, do not make the US’s position clear, their implication is that the United States would only

target cities in the event that their own cities were destroyed.The first nuclear targeting plans existed in the form of SIOP (Single Integrated Operational Plan). This classified document outlined our nuclear policy starting in 1961 until 2004, and now exists in the form of the Operations Plan (OPLAN). The first SIOP specified all out targeting of both military targets and population centers, that is both counterforce *and* countervalue targeting, in both first strike and second strike scenarios. Later SIOPs contained multiple options, including the option to hold the bombing of cities in reserve.

This paper: "The Trump Administration's Nuclear Posture Review (NPR): In Historical Perspective" summarizes how the Kennedy administration began to advocate for a limited war scenario that spared cities:

> *President Kennedy went so far as to endorse Secretary of Defense McNamara's effort to get the Soviets to agree to a "no cities" nuclear targeting rule, which McNamara and the President soon abandoned in the face of objections from NATO and the US Congress as well as the Kremlin that the idea was totally unrealistic. McNamara thereupon did a 180°turn to champion a MAD arms limitation (and retention) pact with the Soviets – to prevent nuclear war by guaranteeing it will be mutually suicidal. The Johnson administration's effort to negotiate such a treaty with Kosygin was aborted in 1968 by the Soviet Union's brutal repression of the reformist Dubcek regime in Czechoslovakia. McNamara continued to work secretly with the military, however, to enlarge the menu in the SIOP (Single Integrated Operational Plan) from which the president could select limited and controlled nuclear responses to a nuclear attack – preserving some possibility of a nuclear cease fire prior to Armageddon.*

Even though McNamara's efforts to change nuclear war plans to spare cities failed, his influence led to changes in the SIOP that for the first time specified a flexible response in nuclear war planning. Nixon would later make additional changes to the SIOP, giving the United States even more flexibility in nuclear targeting scenarios. It is not clear whether the United States ever developed a serious "no cities" strategy during the Cold War, but it did at least lay the foundations for one.

For the first time in US history, President Obama's administration stated the US would not target cities with nuclear weapons. However, this statement did not rule out escalation to countervalue targeting in the midst of a nuclear war, and is best interpreted to mean that the US would only target cities as a retaliatory measure. From the same Historical Perspective paper:

> *Yet Obama, while conceding to this presumed need to be prepared to actually use nuclear weapons in extreme situations, was not about to totally devolve the planning for such use onto the Pentagon…. he was adamant in his guidance to the military that if that crucial threshold ever had to be crossed, all operations had to be "consistent the fundamental principles of the Law of Armed Conflict. Accordingly, plans will … apply the principles of distinction and proportionality and seek to minimize collateral damage to civilian populations and civilian objects. The United States will not intentionally target civilian populations or civilian objects"(US Department of Defense, 2013 US Department of Defense. 2013. Report on Nuclear Employment Strategy of the United States Specified in Section 491 of 10 U.S.C. June 12.*
>
> *The restrictive rules of nuclear engagement were translated into the military's doctrinal language: "The new guidance," elaborated the Pentagon's June 2013*

> *Report on Nuclear Employment Strategy, requires the United States to maintain significant counterforce capabilities [jargon for directed at strategic weapon systems] against potential adversaries. The new guidance does not rely on a 'counter-value' or 'minimum deterrence' strategy [jargon for directed at centers of population]. (US Department of Defense, 2013 US Department of Defense. 2013. Report on Nuclear Employment Strategy of the United States Specified in Section 491 of 10 U.S.C. June 12.*
>
> *Did this mean that the United States was discarding its ultimate assured destruction threat for deterring nuclear war? Clearly not. The guidance was carefully drafted. Does not rely on is different from will not resort to. But more explicitly and openly than previously, the language indicates that assured massive destruction of the enemy country would be the very last resort in an already massively escalating nuclear war, in which all the lesser options had been exhausted and had failed to control the violence.*

President Trump's nuclear policy, as contained in the 2018 Nuclear Posture Review, differs in a number of ways from President Obama's policies, but doesn't substantially change the doctrine of holding the targeting of cities in reserve.

> *If deterrence fails, the initiation and conduct of nuclear operations would adhere to the law of armed conflict and the Uniform Code of Military Justice. The United States will strive to end any conflict and restore deterrence at the lowest level of damage possible for the United States, allies, and partners, and minimize civilian damage to the extent possible consistent with achieving objectives.*
>
> *Every U.S. administration over the past six decades has called for flexible and limited U.S. nuclear response options, in part to support the goal of reestablishing deterrence following its possible failure. This is not because reestablishing deterrence is certain, but because it may be achievable in some cases and contribute to limiting damage, to the extent feasible, to the United States, allies, and partners.*

Conclusion and takeaways:

The US nuclear targeting policy, in so much as public statements and documents reveal, has shifted substantially from a policy of targeting cities by default to a policy that leaves cities as reserve targets for full escalation scenarios. The US policy has never ruled out the possibility of escalation to full countervalue targeting and is unlikely to do so.

The maxim "no plan survives contact with the enemy" is especially worrying from the perspective of nuclear war planning. During the early cold war years described in Daniel Ellsberg's book, the military culture promoted a dedication to nuclear readiness--so much so that officers violated their own protocols to ensure they could launch nuclear weapons in a crisis. Readiness for retaliation, especially full countervalue retaliation, naturally trades off against risk of full escalation.

As both Herman Kahn and Thomas Schelling made clear, communication between legitimate authorities is essential to the ability to negotiate the end of a nuclear conflict. Yet the military value of disabling an enemy's nuclear command, control, and communications (NC3) capabilities is large. This may be the biggest risk to cities; if Moscow and Washington are both destroyed in the early stages of nuclear conflict,

then this could easily escalate to all out countervalue targeting. It is important not only that some command structure with the authority to negotiate remain intact on each side, but also that both parties can communicate with each other and trust that the adversary's command structure is actually intact and capable of negotiation.

Finally, all of this means very little if either of two potential adversaries fail to make plans for a) refraining from initial targeting of cities b) maintaining NC3 capabilities through an initial nuclear strike, and c) have the authority and intention to negotiate a peace & de-escalation. Indeed, public statements by Soviet leadership during the cold war suggested they had no intention of sparing cities in a retaliatory strike, making any possible US policy of gradual escalation potentially useless. Ultimately, it's important to recognize here that not all nuclear war scenarios have equal outcomes, and that both sides in a nuclear conflict could benefit greatly from engaging in strategic restraint.

What is category theory?

Category theory *is* the mathematics of math—specifically, it's a mathematical theory of mathematical structure. It turns out that every kind of mathematics you're likely to encounter in a normal university education is just another kind of category—group theory falls under the category of groups, topology falls under the category of topological spaces, etc.

Specifically, category theory shows that all of the highly diverse mathematical structures that we know of can be broken down into nodes and arrows between nodes. Nodes and arrows define a category in a nutshell—plus something called composition, which basically says that if a bird can fly from Mexico to the USA to Canada, then the bird can also travel "directly" from Mexico to Canada in a way that is equal to the Mexico —> USA —> Canada path. (See "the right to use a name" section.)

Breaking any and every mathematical structure down to nodes and arrows between nodes is a super-general, super-abstract way of studying mathematics. It's like trying to study Shakespeare by breaking down all of his sentences into nouns and verbs. Category theory has a reputation for being painfully abstract—in fact, it's been called "abstract nonsense" by one of its founders. Because of this, it's typically recommended that you have a few mathematical structures under your belt—algebra, groups, topology, etc.—before studying category theory so that you have specific examples to relate the abstractions to. (It would be tough to study verbs if you didn't know about things like running and jumping!)

But while there's only so much to learn about Shakespeare by breaking "to be or not to be" into infinitives, conjunctions, and adverbs, it turns out that the super-general perspective of category theory is incredibly useful in concrete ways. In particular, it turns out that pretty much every cool idea in math is something called an adjoint functor—a special construction that can only be accessed through category theory. A lot of category theorists will tell you that adjoint functors are kind of the *point* of category theory. Adjoints, or adjunction, generalizes *optimization itself*.

Then there is the Yoneda lemma, which is as deep as it is elegant and powerful. We will explore it in depth. (If this series works out.)

You might be wondering what success category theory has found in applications to the sciences. How can you even apply something so general and abstract to our highly specific and concrete reality?

Well, category theory is super general, so whenever you are studying super-general phenomena, it makes sense to think of category theory. What's a super-general phenomenon? For example, the laws of physics! They govern everything, presumably. If you're looking for fundamental rules that apply to everything from tiny particles to enormous planets and the complex living creatures in between, category theory immediately comes to mind.

Then there is biology, which less super-general, unless there really are Martians hiding from our rovers, but organisms have to survive and reproduce under wildly diverse conditions—the planet Earth can throw a lot of stuff at you, from volcanoes to Ice Ages. On some level, organic life clearly has the ability to adapt to all of these conditions—and adapting the same basic thing to lots of different contexts with powerful results is basically what category theory is.

Definitely the biggest applied success for category theory has been in programming. I'd encourage you to look up functional programming, lambda calculus, or just Google something like "programming category theory." It's fascinating, though I'm actually going to deemphasize the programming side of things if anything, as I don't want to distract from the fundamentals.

So what is category theory? Nothing other than the formal generalization of everything. Why should you be interested in it? Because it gives you an incredible bird's-eye view of all of mathematics, and a particular perch, adjunction, that can't be found anywhere else.

This series will be *very* slow paced relative to other introductions—I will *not* assume you know what sets and functions are, to give just one example. If you're comfortable with math or just want to plunge into things a little more, I strongly encourage you to look up the many fantastic introductions to category theory that already exist on the Internet for free as videos, textbooks, blog posts, and papers. This is a series meant for people who either have no exposure to mathematics beyond the high school level or actively want to avoid it! (I'll put it this way: if there was a "Baby Rudin" for category theory, this series would be aiming to be a "Fetal Rudin.")

There's no schedule for these posts, which isn't ideal for learning, but that's just the reality of how this series is going to be made. Coming up is a sequence of posts on the most basic details of defining a category, with an emphasis on developing intuition at each step.

Occam's Razor May Be Sufficient to Infer the Preferences of Irrational Agents: A reply to Armstrong & Mindermann

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

[Epistemic Status: My inside view feels confident, but I've only discussed this with one other person so far, so I won't be surprised if it turns out to be confused.]

Armstrong and Mindermann (A&M) argue "that even with a reasonable simplicity prior/Occam's razor on the set of decompositions, we cannot distinguish between the true decomposition and others that lead to high regret. To address this, we need simple 'normative' assumptions, which cannot be deduced exclusively from observations."

I explain why I think their argument is faulty, concluding that *maybe* Occam's Razor is sufficient to do the job after all.

In what follows I assume the reader is familiar with the paper already or at least with the concepts within it.

Brief summary of A&M's argument:

(This is merely a brief sketch of A&M's argument; I'll engage with it in more detail below. For the full story, read their paper.)

Take a human policy pi = P(R) that we are trying to represent in the planner-reward formalism. R is the human's reward function, which encodes their desires/preferences/values/goals. P() is the human's planner function, which encodes how they take their experiences as input and try to choose outputs that achieve their reward. Pi, then, encodes the overall behavior of the human in question.

Step 1: In any reasonable language, for any plausible policy, you can construct "degenerate" planner-reward pairs that are *almost* as simple as the simplest possible way to generate the policy, yet yield high regret (i.e. have a reward component which is very different from the "true"/"Intended" one.)

- Example: The planner deontologically follows the policy, despite a buddha-like empty utility function
- Example: The planner greedily maximizes the reward function "obedience-to-the-policy."
- Example: Double-negated version of example 2.

It's easy to see that these examples, being constructed from the policy, are *at most* slightly more complex than the simplest possible way to generate the policy, since they could make use of that way.

Step 2: The "intended" planner-reward pair--the one that humans would judge to be a reasonable decomposition of the human policy in question--is likely to be significantly more complex than the simplest possible planner-reward pair.

- Argument: It's really complicated.
- Argument: The pair contains more information than the policy, so it should be more complicated.
- Argument: Philosophers and economists have been trying for years and haven't succeeded yet.

Conclusion: If we use Occam's Razor alone to find planner-reward pairs that fit a particular human's behavior, we'll settle on one of the degenerate ones (or something else entirely) rather than a reasonable one. This could be very dangerous if we are building an AI to maximize the reward.

Methinks the argument proves too much:

My first point is that A&M's argument probably works just as well for other uses of Occam's Razor. In particular it works just as well for the canonical use: finding the Laws and Initial Conditions that describe our universe!

Take a sequence of events we are trying to predict/represent with the lawlike-universe formalism, which posits C (the initial conditions) and then L() the dynamical laws, a function that takes initial conditions and extrapolates everything else from them. L(C) = E, the sequence of events/conditions/world-states we are trying to predict/represent.

Step 1: In any reasonable language, for any plausible sequence of events, we can construct "degenerate" initial condition + laws pairs that are almost as simple as the simplest pair.

- Example: The initial conditions are an empty void, but the laws say "And then the sequence of events that happens is E"
- Example: The initial conditions are simply E, and L() doesn't do anything.

It's easy to see that these examples, being constructed from E, are *at most* slightly more complex than the simplest possible pair, since they could use the simplest pair to generate E.

Step 2: The "intended" initial condition+law pair is likely to be significantly more complex than the simplest pair.

- Argument: It's really complicated.
- Argument: The pair contains more information than the sequence of events, so it should be more complicated.
- Argument: Physicists have been trying for years and haven't succeeded yet.

Conclusion: If we use Occam's Razor alone to find law-condition pairs that fit all the world's events, we'll settle on one of the degenerate ones (or something else entirely) rather than a reasonable one. This could be very dangerous if we are e.g. building an AI to do science for us and answer counterfactual questions like "If we had posted the nuclear launch codes on the Internet, would any nukes have been launched?"

This conclusion may actually be true, but it's a pretty controversial claim and I predict most philosophers of science wouldn't be impressed by this argument for it--even the ones who agree with the conclusion.

Objecting to the three arguments for Step 2

Consider the following hypothesis, which is basically equivalent to the claim A&M are trying to disprove:

Occam Sufficiency Hypothesis: *The "Intended" pair happens to be the simplest way to generate the policy.*

Notice that everything in Step 1 is consistent with this hypothesis. The first degenerate pairs are constructed from the policy, so they are more complicated than the simplest way to generate it, so if that way is via the intended pair, they are more complicated (albeit only slightly) than the intended pair.

Next, notice that the three arguments in support of Step 2 don't really hurt this hypothesis:

Re: first argument: The intended pair can be both very complex and the simplest way to generate the policy; no contradiction there. Indeed that's not even surprising: since the policy is generated by a massive messy neural net in an extremely diverse environment, we should expect it to be complex. What matters for our purposes is *not* how complex the intended pair is, but rather how complex it is *relative to the simplest possible way to generate the policy.* A&M need to argue that the simplest possible way to generate the policy is simpler than the intended pair; arguing that the intended pair is complex is at best only half the argument.

Compare to the case of physics: Sure, the laws of physics are complex. They probably take at least a page of code to write up. And that's aspirational; we haven't even got to that point yet. But that doesn't mean Occam's Razor is insufficient to find the laws of physics.

Re: second argument: The inference from "This pair contains more information than the policy" to "this pair is more complex than the policy" is fallacious. Of course the intended pair contains more information than the policy! All ways of generating the policy contain more information than it. This is because there are many ways (e.g. planner-reward pairs) to get any given policy, and thus specifying any particular way is giving you strictly more information than simply specifying the policy.

Compare to the case of physics: Even once we've been given the complete history of the world (or a complete history of some arbitrarily large set of experiment-events) there will still be additional things left to specify about what the laws and initial conditions truly are. Do the laws contain a double negation in them, for example? Do they have some weird clause that creates infinite energy but only when a certain extremely rare interaction occurs that never in fact occurs? What language are the laws written in, anyway? And what about the initial conditions? Lots of things left to specify that aren't determined by the complete history of the world. Yet this does not mean that the Laws + Initial Conditions are more complex than the complete history of the world, and it certainly doesn't mean we'll be led astray if we believe in the Laws+Conditions pair that is simplest.

Re: third argument: Yes, people have been trying to find planner-reward pairs to explain human behavior for many years, and yes, no one has managed to build a simple algorithm to do it yet. Instead we rely on all sorts of implicit and intuitive heuristics, and we still don't succeed fully. *But all of this can be said about Physics too*. It's not like physicists are literally following the Occam's Razor algorithm--iterating through all possible Law+Condition pairs in order from simplest to most complex and checking each one to see if it outputs a universe consistent with all our observations. And moreover, physicists haven't succeeded fully either. Nevertheless, many of us are still confident that Occam's Razor is in principle sufficient: If we were to follow the algorithm exactly, with enough data and compute, we would eventually settle on a Law+Condition pair that accurately describes reality, and it would be the true pair. Again, maybe we are wrong about that, but the arguments A&M have given so far aren't convincing.

Conclusion

Perhaps Occam's Razor is insufficient after all. (Indeed I suspect as much, for reasons I'll sketch in the appendix) But as far as I can tell, A&M's arguments are at best very weak evidence against the sufficiency of Occam's Razor for inferring human preferences, and moreover they work pretty much just as well against the canonical use of Occam's Razor too.

This is a bold claim, so I won't be surprised if it turns out I was confused. I look forward to hearing people's feedback. Thanks in advance! And thanks especially to Armstrong and Mindermann if they take the time to reply.

Many thanks to Ramana Kumar for hearing me out about this a while ago when we read the paper together.

Appendix: So, is Occam's Razor sufficient or not?

--A priori, we should expect something more like a speed prior to be appropriate for identifying the mechanisms of a finite mind, rather than a pure complexity prior.

--Sure enough, we can think of scenarios in which e.g. a deterministic universe with somewhat simple laws develops consequentialists who run massive simulations including of our universe and then write down Daniel's policy in flaming letters somewhere, such that the algorithm "Run this deterministic universe until you find big flaming letters, then read out that policy" becomes a very simple way to generate Daniel's policy. (This is basically just the "Universal Prior is Malign" idea applied in a new way.)

--So yeah, pure complexity prior is probably not good. But maybe a speed prior would work, or something like it. Or maybe not. I don't know.

--One case that seems useful to me: Suppose we are considering two explanations of someone's behavior: (A) They desire the well-being of the poor, but [insert epicycles here to explain why they aren't donating much, are donating conspicuously, are donating ineffectively] and (B) They desire their peers (and their selves) to *believe* that they desire the well-being of the poor. Thanks to the epicycles in (A), both theories fit the data equally well. But theory B is much more simple. Do we conclude that this person really does desire the well-being of the poor, or not? If we think that even though (A) is more complex it is also more accurate, then yeah it seems like Occam's Razor is insufficient to infer human preferences. But if we instead think "Yeah, this person just really doesn't care, and the proof is how much simpler B is than A" then it seems we really are using something like Occam's Razor to infer human preferences. Of course, this is just one case, so the only way it could prove anything is as a counterexample. To me it doesn't seem like a counterexample to Occam's sufficiency, but I could perhaps be convinced to change my mind about that.

--Also, I'm pretty sure that once we have better theories of the brain and mind, we'll have new concepts and theoretical posits to explain human behavior. (e.g. something something Karl Friston something something free energy?) Thus, the simplest generator of a given human's behavior will probably not divide *automatically* into a planner and a reward; it'll probably have many components and there will be debates about which components the AI should be faithful to (dub these components the reward) and which components the AI should seek to surpass (dub these components the planner.) These debates may be intractable, turning on subjective and/or philosophical considerations. So this is another sense in which I think yeah, definitely Occam's Razor isn't sufficient--for we will also need to have a philosophical debate about what rationality is.

Technical AGI safety research outside AI

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

I think there are many questions whose answers would be useful for technical AGI safety research, but which will probably require expertise outside AI to answer. In this post I list 30 of them, divided into four categories. Feel free to get in touch if you'd like to discuss these questions and why I think they're important in more detail. I personally think that making progress on the ones in the first category is particularly vital, and plausibly tractable for researchers from a wide range of academic backgrounds.

Studying and understanding safety problems

1. How strong are the economic or technological pressures towards building very general AI systems, as opposed to narrow ones? How plausible is the CAIS model of advanced AI capabilities arising from the combination of many narrow services?
2. What are the most compelling arguments for and against discontinuous versus continuous takeoffs? In particular, how should we think about the analogy from human evolution, and the scalability of intelligence with compute?
3. What are the tasks via which narrow AI is most likely to have a destabilising impact on society? What might cyber crime look like when many important jobs have been automated?
4. How plausible are safety concerns about economic dominance by influence-seeking agents, as well as structural loss of control scenarios? Can these be reformulated in terms of standard economic ideas, such as principal-agent problems and the effects of automation?
5. How can we make the concepts of agency and goal-directed behaviour more specific and useful in the context of AI (e.g. building on Dennett's work on the intentional stance)? How do they relate to intelligence and the ability to generalise across widely different domains?
6. What are the strongest arguments that have been made about why advanced AI might pose an existential threat, stated as clearly as possible? How do the different claims relate to each other, and which inferences or assumptions are weakest?

Solving safety problems

1. What techniques used in studying animal brains and behaviour will be most helpful for analysing AI systems and their behaviour, particularly with the goal of rendering them interpretable?
2. What is the most important information about deployed AI that decision-makers will need to track, and how can we create interfaces which communicate this effectively, making it visible and salient?
3. What are the most effective ways to gather huge numbers of human judgments about potential AI behaviour, and how can we ensure that such data is high-quality?

4. How can we empirically test the debate and factored cognition hypotheses? How plausible are the assumptions about the decomposability of cognitive work via language which underlie debate and iterated distillation and amplification?
5. How can we distinguish between AIs helping us better understand what we want and AIs changing what we want (both as individuals and as a civilisation)? How easy is the latter to do; and how easy is it for us to identify?
6. Various questions in decision theory, logical uncertainty and game theory relevant to agent foundations.
7. How can we create secure containment and supervision protocols to use on AI, which are also robust to external interference?
8. What are the best communication channels for conveying goals to AI agents? In particular, which ones are most likely to incentivise optimisation of the goal specified through the channel, rather than modification of the communication channel itself?
9. How closely linked is the human motivational system to our intellectual capabilities - to what extent does the orthogonality thesis apply to human-like brains? What can we learn from the range of variation in human motivational systems (e.g. induced by brain disorders)?
10. What were the features of the human ancestral environment and evolutionary "training process" that contributed the most to our empathy and altruism? What are the analogues of these in our current AI training setups, and how can we increase them?
11. What are the features of our current cultural environments that contribute the most to altruistic and cooperative behaviour, and how can we replicate these while training AI?

Forecasting AI

1. What are the most likely pathways to AGI and the milestones and timelines involved?
2. How do our best systems so far compare to animals and humans, both in terms of performance and in terms of brain size? What do we know from animals about how cognitive abilities scale with brain size, learning time, environmental complexity, etc?
3. What are the economics and logistics of building microchips and datacenters? How will the availability of compute change under different demand scenarios?
4. In what ways is AI usefully analogous or disanalogous to the industrial revolution; electricity; and nuclear weapons?
5. How will the progression of narrow AI shape public and government opinions and narratives towards it, and how will that influence the directions of AI research?
6. Which tasks will there be most economic pressure to automate, and how much money might realistically be involved? What are the biggest social or legal barriers to automation?
7. What are the most salient features of the history of AI, and how should they affect our understanding of the field today?

Meta

1. How can we best grow the field of AI safety? See OpenPhil's notes on the topic.
2. How can spread norms in favour of careful, robust testing and other safety measures in machine learning? What can we learn from other engineering disciplines with strict standards, such as aerospace engineering?
3. How can we create infrastructure to improve our ability to accurately predict future development of AI? What are the bottlenecks facing tools like Foretold.io

and Metaculus, and preventing effective prediction markets from existing?
4. How can we best increase communication and coordination within the AI safety community? What are the major constraints that safety faces on sharing information (in particular ones which other fields don't face), and how can we overcome them?
5. What norms and institutions should the field of AI safety import from other disciplines? Are there predictable problems that we will face as a research community, or systemic biases which are making us overlook things?
6. What are the biggest disagreements between safety researchers? What's the distribution of opinions, and what are the key cruxes?

Particular thanks to Beth Barnes and a discussion group at the CHAI retreat for helping me compile this list.

Learning from other people's experiences/mistakes

One of the fastest way to learn is to learn from someone else's mistakes and experiences. This short-cuts a lot of unnecessary trial and error and can save significant time. However, one is sorely tempted to repeat the experiences/mistakes of others. One may think, that they are smarter/luckier than the others who made those mistakes. One may not trust that the right lessons were learnt by others in their experiences. One may think there is loss of agency in just following along a path someone else prescribed. What are some guidelines do you use to learn from others experiences? How do you judge their lessons are worth following? How do you stop yourself from attempting to make those mistakes yourself.

Prospecting for Conceptual Holes

Imagine you wanted to explore the number line. You start at 1 and then you explore 2, 3, 4 and so on. No matter how long you do this you'll never discover −1, ½, 2π, ∞, i, or a single noncomputable number. That's because the most interesting numbers are precisely those located in directions conceptually orthogonal to whatever numbers you've already explored.

The easiest concepts to learn are the ideas ahead of you on a path you're already taking. If you know classical mechanics then it's straightforward to learn relativity. You could crack turbulence without ever changing fields.

The second-easiest things to learn are those outside your specialization but inside of your culture's intellectual tradition. If you grew up in the West then you know there are such things as physics and painting even if you never learned how to do them.

The second-hardest things to learn are those outside your society's intellectual tradition. Consider for example *kenshō*[1]. It's possible to live an entire lifetime outside of Japan without ever discovering that such a concept exists. Westerners don't choose not to experience *kenshō*. This choice is made invisibly.

The hardest things to learn are those belonging to domains of knowledge which haven't been (openly) conceptualized anywhere, such as photography before the invention of the camera. This category includes zero-day exploits.

We can number each kind of conceptual hole from least to most orthogonal.

- Type 1: Concepts you are aware of but do not understand.
- Type 2: Concepts you are not aware of, but which belong to a field you are aware of.
- Type 3: Fields of knowledge you do not know exist, but which exist.
- Type 4: Fields of knowledge that remain genuinely secret or have not been invented.

Expertise is about mastering one field, but general intelligence is how prepared you are for something new. The more different kinds of concepts you understand the better you will be at solving new kinds of problems. This is behind the principle of learning that breadth of knowledge equals breadth of transference.

Filling holes of Type 1 is how you build expertise in a field. Holes of Types 2-4 are better for building broader (and therefore more transferable) knowledge. In other words, holes of greater-numbered types are better for increasing your general intelligence. If this is your goal then holes of Type 2 and 3 are the most valuable.

I find holes of Type 3 to be so valuable that just knowing where they are improves my creativity at solving problems. (It also turns them into holes of Type 2.) You can find holes of Type 3 systematically by mastering the language of a sufficiently foreign culture. I know this works with written Chinese and spoken Pirahã. I hypothesize it's also true of Arabic, ASL, Korean and the Khoisan languages.

Filling holes of Type 2 is almost straightforward. Pick anything you're bad at that lots of other people do and develop a basic competence. This turns holes of Type 2 into holes of Type 1.

Holes of Type 4 can be individually very valuable. That is, a single hole of Type 4 can earn you lots of power and money. But holes of Type 4 are so hard to find, verify and exploit that you can't build a broad base of knowledge out of them.

I like collecting conceptual holes, especially those of Types 2 and 3. They're so interesting. Whenever I discover a new hole it opens up an entire tree of knowledge orthogonal to everything I used to know.

---

1. *Kenshō* (Japanese kanji: 見性) is a subjective state of mind associated with Zen meditation. ↩

Ideal Number of Parents

I'll often hear people say with varying levels of seriousness that the ideal number of parents to have is something large, like maybe five. Children, especially infants, can be an enormous amount of work, which is certainly easier spread out among more people. Kids also like getting a lot of attention, can require a lot of money, and, since different adults are good with kids in different ways, can benefit from having a range of adults in their lives. But while having many people involved in taking care of the kids is great, I'm not sure having all of them be co-equal parents is a good approach.

First, a question: why are people excited about putting so much of themselves into parenting, when if you wanted to pay for similar levels of childcare it would be incredibly expensive? Not to suggest that parenting is simply unpaid childcare; thinking about what why these superficially similar situations lead to such different levels of desire helps illuminate what's important about parenting.

Answers will, of course, vary based on individual perspectives and drives. For some people the answer is "I'm not excited about this, which is why I don't want kids", but for people who do want to be a parent I think it's common for things to trace back through two points:

- Having a substantial say in how the child is raised.
- Knowing that this is a life-long relationship.

For the first point, the more people you have co-parenting the less say each one has and the harder it is to reach agreement. Parents can have different ideas on what is safe, how discipline should work, how much help to give, how to do food, value of different kinds of toys/screens/games, co-sleeping, night training, potty training, is it ok to microwave baby milk, what rules to have for sharing, how structured the day should be, when they're ready to go outside alone, how to do money, what to do for childcare, when bedtime should be, what's important in schooling, how important is predictability, how to handle various unique challenges most kids have in some form, how to do presents, when to let them try a thing, what medical treatments make sense, how much to let them make their own decisions, whether to let them ask people for things when it's kind of rude, how much to push them, when to encourage an interest, how to build responsibility, and how to balance all kinds of tricky tradeoffs.

For important issues having more people involved in the decision could make it more likely you are to get good decisions, but you need to balance this against how hard it is to get people to come to agreement on things they feel very strongly about. Unless all the parents have an incredibly close sense of how kids should be raised there will be a lot of these, based on different childhood experience, different parenting philosophy, how to weigh different factors, etc. This is hard enough with two people, and it seems like something that gets substantially more difficult the more parents there are.

For the second point, the permanent nature of the relationship allows a kind of parent-child bonding that people are understandably wary of in more temporary arrangements. I care enormously about what happens to my kids, and part of that is knowing that they're my responsibility no matter what. Getting this kind of assurance of permanence with a larger number of parents is legally somewhere between "very difficult" and "not possible" in a society where only some parents will have official

status. The legal parent(s) could at some point, if things fall apart, cut the others out. If you think co-parents wouldn't do this, consider how many loving relationships collapse into spitefests in divorce. Even if we fixed the legal aspect, however, the more people you have in parental roles the more likely there is to be some kind of falling-out over the years, and joint custody among large numbers of households wouldn't work well.

Other aspects that could be a problem, however, seem like they could be managed with good communication, good culture, and dividing things. For example, I do the kids breakfast and pack Lily's lunch in the morning (hence the thermos experimentation). I then pay attention to what comes home uneaten in the lunchbox to try to figure out what I should send next time. In figuring out what to send I also pay attention to what she's been eating and not eating at breakfast and dinner. [1] Even if we had several other co-parents we could still divide things up so this was all on one person and avoid having to manage this process across the full number of parents. I do think the ways parenting work tends to drift towards the people who are already doing the most, because they're currently best at it, are more of an issue, but a surmountable one if you're attentive.

Overall I do think something with more parents can work, and I'm excited people are trying out new approaches. I think it could turn out to be really positive for children to have so many adults strongly invested in their well-being. But I think children having one or two parents in a strong and stable community/household of family/friends probably works better than a larger number of fully-equal parents.

[1] I initially tried an approach of asking her each day what she wanted for lunch, but it turns out she's pretty bad at predicting what she's going to want to eat. So my current strategy is that I pack what I think she'll eat, and then if she wants me to pack something in addition she can ask me to and I'll do that as well. I do eventually want to move to her packing her own lunch and getting good at figuring out what she wants to eat, but at least for now it's much more important to me that she be getting enough to eat.

Comment via: facebook

The Missing Piece

In computer science, compiler is a program that translates things that programmers write, such as "if CloseButtonPressed then CloseApplication", to things that computers are able to understand and execute, such as "00010101110010101110110001111000101010111110".

When the first compiler was written, someone had to write it in the machine code, the zeroes and ones that computers understand. They had to type in the numbers, or, more likely, punch holes into punchcards.

But from that point on, programmers were able to write human-readable programs instead of the inscrutable machine code. They just had to use the existing compiler to translate from the former to the latter.

Now imagine you are tasked with writing the second, improved, version of the compiler.

You can, of course, do the same thing again. You can type in the numbers, or maybe punch zeroes and ones into punchcards. But writing machine code is hard.

And now you have a better option at your disposal: You can write the new version of the compiler in a human-readable language and use the old version of the compiler to translate it into machine code!

Similarly, the third version of the compiler can be compiled by the second version of the compiler, the fourth by the third and so on.

That, by the way is not a hypothetical scenario. That's how the compilers are written nowadays.

Now, here comes the question: What if all the software disappeared overnight? What if the only thing we were left with was the human-readable source code of the latest version of the compiler? Would we be able to reconstruct the language and the compiler and get everything running again?

On one hand, the source code contains all the relevant information. It contains the full description of the language, its syntax and its semantics. It contains all the instructions needed to translate it into the machine code. It should be therefore possible to use it to create a functional compiler again.

But on the other hand, how exactly would you do that? The source code is, after all, just a pile of text. And there's no compiler to turn it into the machine code, into something that computer understands. It sounds like you are faced with a kind of chicken-and-egg problem.

Clearly, the source code is not enough. There's some information missing. But it's hard to pinpoint what exactly the missing part may be.

***

What if we found an undamaged T-Rex DNA in the gut of ancient mosquito trapped inside a piece of amber? Would we be able to grow an adult T-Rex? Would we be able

to find out what colour they were, whether they had feathers and whether males attracted females by nightingale-like song?

On one hand, one regularly reads news about scientists being on the verge of resurrecting the wooly mammoth. It am no embryologist, but as far as I understand, the idea is that the DNA reconstructed from the frozen remains of mammoths found in Siberia would be injected into an elephant egg and grown either in an elephant female or maybe ex-vivo.

On the other hand, dinosaurs have no close living relatives. For T-Rex, there's no equivalent of what elephant is to mammoth. Maybe the DNA could be injected into an ostrich egg, the birds being the closest surviving relatives to dinosaurs. But given the evolutionary distance between T-Rex and ostrich, how likely it is that the biochemistry and embryology of the ostrich egg is similar enough to the biochemistry and embryology of the T-Rex egg to produce a viable offspring? I wouldn't bet much on it.

And even if that was possible, what if there was no living organism left, just a piece of DNA? Would we still be able to reconstruct the animal?

Although DNA encodes all the information about the organism, there's clearly something missing. And, again, it's hard to put your finger on what exactly it may be.

***

Central African economy is failing badly. CAR authorities therefore look around for inspiration and decide to adopt the Swiss model. Switzerland, after all, is a small multi-ethnic, multi-confessional country, just like CAR, and it was able to raise from being an obscure and poor backwater country to one of the world's most wealthy countries in just few decades.

CAR therefore adopts the Swiss constitution, Swiss political system, Swiss law, Swiss approach to bank secrecy, Swiss educational system. They establish a church choir in every village and start eating fondue.

But to everyone's surprise nothing changes. CAR economy is doing as badly as ever. Corruption is rife. The rule of law falters. The institutions that run like clockwork in Switzerland fail mysteriously in CAR.

Clearly, there's something about Swiss society that wasn't captured in the laws and institutions introduced in CAR. But, once again, it's hard to say what the missing piece could possibly be.

October 27th, 2019

>

Maybe Lying Doesn't Exist

In "Against Lie Inflation", the immortal Scott Alexander argues that the word "lie" should be reserved for knowingly-made false statements, and not used in an expanded sense that includes unconscious motivated reasoning. Alexander argues that the expanded sense draws the category boundaries of "lying" too widely in a way that would make the word less useful. The hypothesis that predicts everything predicts nothing: in order for "Kevin lied" to *mean something*, some possible states-of-affairs need to be identified as *not* lying, so that the statement "Kevin lied" can correspond to redistributing conserved probability mass *away from* "not lying" states-of-affairs *onto* "lying" states-of-affairs.

All of this is entirely correct. But Jessica Taylor (whose post "The AI Timelines Scam" inspired "Against Lie Inflation") wasn't arguing that *everything* is lying; she was just using a *more* permissive conception of lying than the one Alexander prefers, such that Alexander didn't think that Taylor's definition could stably and consistently identify non-lies.

Concerning Alexander's arguments against the expanded definition, I find I have one strong objection (that appeal-to-consequences is an invalid form of reasoning for optimal-categorization questions for essentially the same reason as it is for questions of simple fact), and one more speculative objection (that our intuitive "folk theory" of lying may actually be empirically mistaken). Let me explain.

(A small clarification: for myself, I notice that I *also* tend to frown on the expanded sense of "lying". But the *reasons* for frowning matter! People who superficially agree on a conclusion but for *different reasons*, are not really on the same page!)

Appeals to Consequences Are Invalid

> There is no method of reasoning more common, and yet none more blamable, than, in philosophical disputes, to endeavor the refutation of any hypothesis, by a pretense of its dangerous consequences[.]
>
> —David Hume

Alexander contrasts the imagined consequences of the expanded definition of "lying" becoming more widely accepted, to a world that uses the restricted definition:

> [E]veryone is much angrier. In the restricted-definition world, a few people write posts suggesting that there may be biases affecting the situation. In the expanded-definition world, those same people write posts accusing the other side of being liars perpetrating a fraud. I am willing to listen to people suggesting I might be biased, but if someone calls me a liar I'm going to be pretty angry and go into defensive mode. I'll be less likely to hear them out and adjust my beliefs, and more likely to try to attack them.

But this is an appeal to consequences. Appeals to consequences are invalid because they represent a map–territory confusion, an attempt to optimize our *description* of reality at the expense of our ability to describe reality *accurately* (which we need in order to *actually* optimize reality).

(Again, the appeal is still invalid even if the conclusion—in this case, that unconscious rationalization shouldn't count as "lying"—might be true for *other reasons*.)

Some aspiring epistemic rationalists like to call this the "Litany of Tarski". *If* Elijah is lying (with respect to whatever the optimal category boundary for "lying" turns out to be according to our standard Bayesian philosophy of language), *then* I desire to believe that Elijah is lying (with respect to the optimal category boundary according to ... *&c.*). *If* Elijah is *not* lying (with respect to ... *&c.*), *then* I desire to believe that Elijah is *not* lying.

If the one comes to me and says, "Elijah is not lying; to support this claim, I offer this-and-such evidence of his sincerity," then this is right and proper, and I am eager to examine the evidence presented.

If the one comes to me and says, "You should choose to define *lying* such that Elijah is not lying, because if you said that he was lying, then he might feel angry and defensive," this is *insane*. The map is not the territory! If Elijah's behavior is, *in fact*, deceptive—if he says things that cause people who trust him to be worse at anticipating their experiences when he reasonably could have avoided this—I can't make his behavior not-deceptive by *changing the meanings of words*.

Now, I *agree* that it might very well empirically be the case that if I *say* that Elijah is lying (where Elijah can hear me), he might get angry and defensive, which could have a variety of negative social consequences. But that's not an argument for changing the definition of lying; that's an argument that I have an incentive to lie about whether I think Elijah is lying! (Though Glomarizing about whether I think he's lying might be an even better play.)

Alexander is concerned that people might strategically equivocate between different definitions of "lying" as an unjust social attack against the innocent, using the classic motte-and-bailey maneuver: first, argue that someone is "lying (expanded definition)" (the motte), then switch to treating them as if they were guilty of "lying (restricted definition)" (the bailey) and hope no one notices.

So, I agree that this is a very real problem. But it's worth noting that the problem of equivocation between different category boundaries associated with the same word applies *symmetrically*: if it's possible to use an expanded definition of a socially-disapproved category as the motte and a restricted definition as the bailey in an unjust attack against the innocent, then it's *also* possible to use an expanded definition as the bailey and a restricted definition as the motte in an unjust defense of the guilty. Alexander writes:

> The whole reason that rebranding lesser sins as "lying" is tempting is because everyone knows "lying" refers to something very bad.

Right—and conversely, because everyone knows that "lying" refers to something very bad, it's tempting to rebrand lies as lesser sins. Ruby Bloom explains what this looks like in the wild:

> I worked in a workplace where lying was commonplace, conscious, and system 2. Clients asking if we could do something were told "yes, we've already got that feature (we hadn't) and we already have several clients successfully using that (we hadn't)." Others were invited to be part an "existing beta program" *alongside others just like them* (in fact, they would have been the very first). When I objected, I was told "no one wants to be the first, so you have to say that."

> [...] I think they lie to themselves that they're not lying (so that if you search their thoughts, they never think "I'm lying")[.]

If your interest in the philosophy of language is primarily to *avoid being blamed for things*—perhaps because you perceive that you live in a Hobbesian dystopia where the primary function of words is to elicit actions, where the denotative structure of language was eroded by political processes long ago, and all that's left is a standardized list of approved attacks—in that case, it makes perfect sense to worry about "lie inflation" but not about "lie deflation." If describing something as "lying" is primarily a weapon, then applying extra scrutiny to uses of that weapon is a wise arms-restriction treaty.

But if your interest in the philosophy of language is to improve and refine the uniquely human power of vibratory telepathy—to construct shared maps that reflect the territory—if you're interested in revealing what kinds of deception are *actually happening*, and why—

(in short, if you are an aspiring epistemic rationalist)

—then the asymmetrical fear of false-positive identifications of "lying" but not false-negatives—along with the focus on "bad actors", "stigmatization", "attacks", *&c.*—just looks *weird*. What does *that* have to do with maximizing the probability you assign to the right answer??

The Optimal Categorization Depends on the Actual Psychology of Deception

> *Deception*
> *My life seems like it's nothing but*
> *Deception*
> *A big charade*
>
> *I never meant to lie to you*
> *I swear it*
> *I never meant to play those games*
>
> —"Deception" by Jem and the Holograms

Even if the fear of rhetorical warfare isn't a legitimate reason to avoid calling things lies (at least privately), we're still left with the main objection that "lying" is a *different thing* from "rationalizing" or "being biased". Everyone is biased in some way or another, but to *lie* is "[t]o give false information intentionally with intent to deceive." Sometimes it might make sense to use the word "lie" in a noncentral sense, as when we speak of "lying to oneself" or say "Oops, I lied" in reaction to being corrected. But it's important that these senses be explicitly acknowledged as noncentral and not conflated with the central case of knowingly speaking falsehood with intent to deceive—as Alexander says, conflating the two can only be to the benefit of *actual liars*.

Why would anyone disagree with this obvious ordinary view, if they *weren't* trying to get away with the sneaky motte-and-bailey social attack that Alexander is so worried about?

Perhaps because the ordinary view relies an implied theory of human psychology that we have reason to believe is false? What if *conscious* intent to deceive is typically

absent in the most common cases of people saying things that (they would be capable of realizing upon being pressed) they know not to be true? Alexander writes—

> So how will people decide where to draw the line [if egregious motivated reasoning can count as "lying"]? My guess is: in a place drawn by bias and motivated reasoning, same way they decide everything else. The outgroup will be lying liars, and the ingroup will be decent people with ordinary human failings.

But if the word "lying" is to actually *mean something* rather than just being a weapon, then the ingroup and the outgroup *can't both be right*. If symmetry considerations make us doubt that one group is really that much more honest than the other, that would seem to imply that *either* both groups are composed of decent people with ordinary human failings, *or* that both groups are composed of lying liars. The first description certainly *sounds nicer*, but as aspiring epistemic rationalists, we're *not allowed to care* about which descriptions sound nice; we're *only* allowed to care about which descriptions match reality.

And if all of the concepts available to us in our native language fail to match reality in different ways, then we have a tough problem that may require us to innovate.

The philosopher Roderick T. Long writes—

> Suppose I were to invent a new word, "zaxlebax," and define it as "a metallic sphere, like the Washington Monument." That's the definition—"a metallic sphere, like the Washington Monument." In short, I build my ill-chosen example into the definition. Now some linguistic subgroup might start using the term "zaxlebax" as though it just meant "metallic sphere," or as though it just meant "something of the same kind as the Washington Monument." And that's fine. But my definition incorporates both, and thus conceals the false assumption that the Washington Monument is a metallic sphere; any attempt to use the term "zaxlebax," meaning what I mean by it, involves the user in this false assumption.

If self-deception is as ubiquitous in human life as authors such as Robin Hanson argue (and if you're reading this blog, this should not be a new idea to you!), then the ordinary concept of "lying" may actually be analogous to Long's "zaxlebax": the standard *intensional* definition ("speaking falsehood with conscious intent to deceive"/"a metallic sphere") fails to match the most common extensional examples that we want to use the word for ("people motivatedly saying convenient things without bothering to check whether they're true"/"the Washington Monument").

Arguing for this *empirical* thesis about human psychology is beyond the scope of this post. But *if* we live in a sufficiently Hansonian world where the *ordinary* meaning of "lying" fails to carve reality at the joints, then authors are faced with a tough choice: either be involved in the false assumptions of the standard believed-to-be-central intensional definition, or be deprived of the use of common expressive vocabulary. As Ben Hoffman points out in the comments to "Against Lie Inflation", an earlier Scott Alexander didn't seem shy about calling people liars in his classic 2014 post "In Favor of Niceness, Community, and Civilization"—

> Politicians lie, but not *too much*. Take the top story on Politifact Fact Check today. Some Republican claimed his supposedly-maverick Democratic opponent actually voted with Obama's economic policies 97 percent of the time. Fact Check explains that the statistic used was actually for all votes, not just economic votes, and that members of Congress typically have to have >90% agreement with their president because of the way partisan politics work. **So it's a lie, and is**

> **properly listed as one.** [bolding mine —ZMD] But it's a lie based on slightly misinterpreting a real statistic. He didn't just totally make up a number. He didn't even just make up something else, like "My opponent personally helped design most of Obama's legislation".

Was the politician consciously lying? Or did he (or his staffer) arrive at the misinterpretation via unconscious motivated reasoning and then just not bother to scrupulously check whether the interpretation was true? And how could Alexander know?

Given my current beliefs about the psychology of deception, I find myself inclined to reach for words like "motivated", "misleading", "distorted", *&c.*, and am more likely to frown at uses of "lie", "fraud", "scam", *&c.* where intent is hard to establish. But even while frowning internally, I want to avoid tone-policing people whose word-choice procedures are calibrated differently from mine when I think I understand the structure-in-the-world they're trying to point to. Insisting on replacing the six instances of the phrase "malicious lies" in "Niceness, Community, and Civilization" with "maliciously-motivated false belief" would just be *worse writing*.

And I *definitely* don't want to excuse motivated reasoning as a mere ordinary human failing for which someone can't be blamed! One of the key features that distinguishes motivated reasoning from simple mistakes is the way that the former *responds to incentives* (such as being blamed). If the elephant in your brain thinks it can *get away with lying* just by keeping conscious-you in the dark, it should think again!

Epistemic Spot Checks: The Fall of Rome

Introduction

Epistemic spot checks are a series in which I select claims from the first few chapters of a book and investigate them for accuracy, to determine if a book is worth my time. This month's subject is *The Fall of Rome*, by Bryan Ward-Perkins, which advocates for the view that Rome fell, and it was probably a military problem.

Like August's *The Fate of Rome*, this spot check was done as part of a collaboration with Parallel Forecasting and Foretold, which means that instead of resolving a claim as true or false, I give a confidence distribution of what I think I would answer if I spent 10 hours on the question (in reality I spent 10-45 minutes per question). Sometimes the claim is a question with a numerical answer, sometimes it is just a statement and I state how likely I think the statement is to be true.

This spot check is subject to the same constraints as *The Fate of Rome*, including:

1. Some of my answers include research from the forecasters, not just my own.
2. Due to our procedure for choosing questions, I didn't investigate all the claims I would have liked to.

Claims

Claim made by the text: *"[Emperor Valerian] spent the final years of his life as a captive at the Persian Court"*
Question I answered*: what is the chance that is true?*
My answer*: I estimate a chance of (99 – 3*lognormal(0,1)) that Emperor Valerian was captured by the Persians and spent multiple years as a prisoner before dying in captivity.*

You don't even have to click on the Wikipedia page to confirm this is the common story: it's in the google preview for "emperor valerian". So the only question is the chance that all of history got this wrong. Wikipedia lists five primary sources, of which I verified three. https://www.ancient-origins.net/history/what-really-happened-valerian-was-roman-emperor-humiliated-and-skinned-hands-enemy-008598 raises questions about how badly Valerian was treated, but not that he was captive.

My only qualm is the chance that this could be a lie perpetuated at the time. Maybe Valerian died and the Persians used a double, maybe something weirder happened. System 2 says the chance of this is < 10% but gut says < 15%.

Claim made by the text: *"What had totally disappeared, however, were the good-quality, low-value items, made in bulk, and available so widely in the Roman period"*
Question I answered*: What is the chance mass-produced, low-value items available so widely in the Roman period, disappear in Britain by 600 AD?*

***My answer**: I estimate a chance of (64 to 93, normal distribution) that mass-produced, low-value items were available in Britain during Roman rule and not after 600 AD.*

This was one of the hardest claims to investigate, because it represents original research by Ward-Perkins. I had basically given up on answering this without a pottery PhD until google suggestions gave me the perfect article.

This is actually a compound claim by Ward-Perkins:

1. Roman coinage and mass-produced, low-cost, high-quality pottery disappeared from Britain and then the rest of post-Roman Europe.
2. The state of pottery and coinage is a good proxy for the state of goods and trades as a whole, because they preserve so amazingly well and are relatively easy to date.

Data points:

- https://brewminate.com/the-perils-of-periodization-roman-ceramics-in-britain-after-400-ce/ Cites exactly the pattern Ward-Perkins describes in pottery and coins, citing Ward-Perkins and 4 other source I couldn't verify. This seems like very strong confirmation of W-P's hypothesis. However it leaves open the chance that this is an area of contention, of which W-P and brewminate happen to be on the same side. Brewminate's main focus is not on the British empire, but on pottery in the gap between the Roman's and the Anglo-Saxons. I estimate that would bias them towards increasing the size of the discontinuity between Roman and post-Roman Britain.
- https://www.cambridge.org/core/journals/britannia/article/romanization-of-pottery-assemblages-in-the-east-and-northeast-of-england-during-the-first-century-ad-a-comparative-analysis/AE110A5FE5018C87E8DB284C458F1CC8 (cited by brewminate) establishes that pottery was romanized and centralized in the 1st century AD
- Roman Pottery by Kevin Greene establishes approximately the same thing, and was found independently of brewminate.
- https://www.romanobritain.org/10_crafts/rca_roman_british_pottery.php claims most pottery in Britain was imported even before Rome invaded (43 AD)
- https://www.romanobritain.org/10_crafts/rca_roman_british_pottery.php claims during Roman rule Britain had several high quality pottery production centers
- https://www.thegreatcoursesdaily.com/britain-after-the-romans-left/ (I think a very mainstream source) repeats the story that pottery is informative and pottery shows a decline after 400s.
- Searching for "the fall of rome ward-perkins criticism" turns up nothing interesting in the first two pages.
- Searched for "late antiquity british pottery" found nothing. "late antiquity" is the term for the narrative that Rome didn't fall, it just transformed, and is what Ward-Perkins is directly arguing against, so I'd expect criticism of him to be coded with that term.
- https://brill.com/view/book/edcoll/9789004309784/B9789004309784-s016.xml says "Inter-regional trade was not a motor for prosperity in the later Roman period in Britain"
-
    - Focuses on how amphorae were never really abundant in Britain
    - Chart stops at 400 AD

    - Graph showing large drops in amphorae distribution by 410 AD

- https://www.academia.edu/1353493/Form_function_and_technology_in_pottery_production_from_Late_Antiquity_to_the_early_Middle_Ages uses term “late antiquity” and cites end of large scale pottery production in Rome

If we believe Ward-Perkins and Brewminate, I estimate the chances that pottery massively declined at 95-99, times 80-95 that other good declined with them. There remains the chances that the historical record is massively misleading (very unlikely with pots, although I don’t know how likely it is to have missed sites entirely), and that W-P et al are misinterpreting the record. I would be very surprised if so many sites had been missed as to invalidate this data, call it 5-15%. Gut feeling, 5-20% chance the W-P crowd are exaggerating the data, but given the absence of challenges, not higher than that and not a significant chance they’re just making shit up.

(95 to 99)*(85 to 95) * (80 to 95) = 64 to 93%

Claim made by the text: *The Romans had mass literacy, which declined during the Dark Ages.*
Question I answered*: “[% population able to read at American 1st grade level during Imperial Rome] – [% population able to do same in the same geographic area in 1000 AD] = N%. What is N?”*
My answer*: I estimate that there is a 95% chance [Roman literacy] – [Dark Ages literacy] = (0 to 60, normal distribution)*

Data Points:

- https://en.wikipedia.org/wiki/Roman_Empire#Literacy,_books,_and_education
- “Estimates of the average literacy rate in the Empire range from 5 to 30% or higher, depending in part on the definition of “literacy” I’ll use the high end, since I’m using a pretty minimal definition of literacy
- https://www.quora.com/What-were-literacy-rates-in-Medieval-Europe-How-did-they-compare-to-literacy-rates-in-the-Roman-Empire
- “literacy levels were probably much lower than in Roman times during most of the Middle Ages – perhaps only 6% in England in 1300”
- https://www.quora.com/What-was-the-level-of-literacy-among-the-people-of-the-Roman-Empire-first-to-third-centuries-AD-If-only-the-wealthy-were-literate-what-was-the-point-of-the-written-propaganda-on-the-ancient-monuments claims literacy was low, 5-10 % of the population, with a maximum of 20 per cent.
- http://trisagionseraph.tripod.com/literacyf.html is hideous and should die

The highest estimate of literacy in Roman Empire I found is 30%. Call it twice that for ability to read at a 1st grade level in cities. So the range is 5%-60%.

The absolute lowest the European 1000AD literacy rate could be is 0; the highest estimate is 5% (and that was in the 1300s, which were probably more literate). From the absence of graffiti I infer that even minimal literacy achievement dropped a great deal.

Maximum = 60%-1% = 59%
Minimum = 5%-5%=0

Claim made by the text: *"What some people describe as "the invasion of Rome by Germanic barbarians", Walter Goffart describes as "the Romans incorporating the Germanic tribes into their citizenry and setting them up as rulers who reported to the empire." and "Rome did fall, but only because it had voluntarily delegated its own power, not because it had been successfully invaded"."*
Question I answered: *What is my confidence that this accurately represents historian Walter Goffart's views?*
My answer: *I estimate that after 10 hours of research, I would be 68-92% confident this describes Goffart's views accurately.*

Data points:

- https://blog.oup.com/2005/12/the_fall_of_rom/
  - Peter Heather: The most influential statement of this, perhaps, is Walter Goffart's brilliant aphorism that the fall of the Western Empire was just 'an imaginative experiment that got a little out of hand'. Goffart means that changes in Roman policy towards the barbarians led to the emergence of the successor states, dependant on barbarian military power and incorporating Roman institutions, and that the process which brought this out was not a particularly violent one.
- https://www.goodreads.com/book/show/1680215.Barbarians_and_Romans_A_D_418_584?from_search=true
  - "Despite intermittent turbulence and destruction, much of the Roman West came under barbarian control in an orderly fashion."
- https://press.princeton.edu/titles/1036.html
  - "Despite intermittent turbulence and destruction, much of the Roman West came under barbarian control in an orderly fashion. Goths, Burgundians, and other aliens were accommodated within the provinces without disrupting the settled population or overturning the patterns of landownership. Walter Goffart examines these arrangements and shows that they were based on the procedures of Roman taxation, rather than on those of military billeting (the so-called *hospitalitas* system), as has long been thought. Resident proprietors could be left in undisturbed possession of their lands because the proceeds of taxation,rather than land itself, were awarded to the barbarian troops and their leaders."
- https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1478-0542.2008.00523.x
  - "the barbarians and Rome, instead of being in constant conflict with each other, occupied a joined space, a single world in which both were entitled to share. What we call the barbarian invasions was primarily a drawing of foreigners into Roman service, a process sponsored, encouraged, and rewarded by Rome. Simultaneously, the Romans energetically upheld their supremacy. Many barbarian peoples were suppressed and vanished; the survivors were persuaded and learned to shoulder Roman tasks. Rome was never discredited or repudiated. The future endorsed and carried forward what the Empire stood for in religion, law, administration, literacy, and language."
- https://books.google.com/books/about/Rome_s_Fall_and_After.html?id=55pDIwvWnpoC "Rome's Fall and After" indicates Goffat does believe Rome fell. But suggests its main problem was constantinople, not interactions with barbarians at all. So top percentage correct = 90%)

This seems pretty conclusive that Goffart thought Barbarians were accommodated rather than conquered the area (so my minimum estimate that the summary was correct must be greater than 50%). However it's not clear how much power he thought they took, or whether rome fell at all. This could be a poor restatement, or it could be that if I read Goffart's actual work and not just book jacket blurbs I'd agree.

Question I answered: *Chance Elizabeth would recommend this book as a reliable source on the topic to an interested friend, if they asked tomorrow (8/31/19)?*
My answer: *There is a (91-99%, normal distribution) chance I would recommend this to a friend.*

99% is in range, because I definitely think it's worth reading if they're interested in the topic. I think I'd recommend it before Fate of Rome, because it establishes that rome fell more concretely.

Is there a chance I wouldn't recommend it?

- They could have already read it
- They could be more interested in disease and climate change (in which case I'd recommend Fate)
- I could forget about it
- I could not want to take responsibility for their reading.
- I could be unconfident that Fall was better than what they'd find by chance.
    - This feels like the biggest one.
    - But the question doesn't say "best book", it just says "reliable source"
    - Only real qualm on that is that is normal history book qualms

So the minimum is 91%

Bonus Claims

These are the claims I didn't check, but other people made predictions on how I would guess. Note that at this point the predictions haven't been very accurate- whether they're net positive depends on how you weight the questions. And Foretold is beta software that hasn't prioritized export yet, so I'm using *shudder* screen shots. But for the sake of completeness:

Claim made by the text: The Fall of Rome: Roman Pottery pre-400AD was high quality and uniform.
Predicted answer: 29.9% to 63.5% chance this claim is correct

Claim made by the text: "In Britain new coins ceased to reach the island, except in tiny quantities, at the beginning of the fifth century"
Predicted answer: 31.6% to 94% chance this claim is correct

Claim made by the text: The Fall of Rome: [average German soldiers' height] – [average Roman soldiers' height] = N feet. What is N? .

Predicted answer: -0.107 to 0.61 ft.

Claim made by the text: The Romans chose to cede local control of Gaul to the Germanic tribes in the 400s, as opposed to losing them in a military conquest.
Predicted answer: 28.5% to 85.6% chance this claim is correct

Claim made by the text: The Germanic tribes who took over local control of Gaul in the 400s reported to the Emperor.
Predicted answer: 4.77% to 50.9% chance this claim is correct

Conclusion

The Fall of Rome did very well on spot-checking- no outright disagreements at all, just some uncertainties.

On the other hand, *The Fall of Rome* barely mentions disease and doesn't mention climate change at all, which my previous book, *The Fate of Rome*, claimed to be the main causes of the fall. *The Fate of Rome* did almost as well in epistemic spot checking as *Fall*, yet they can't both be correct. What's going on? I'm going to address that in a separate post, because I want to be able to link to it without forcing people to read this entire spot check.

In terms of readability, *Fall* starts slowly but the second half is by far the most interested I have ever been in pottery or archeology.

[Many thanks to my Patreon patrons and Parallel Forecast for financial support for this post]

Does combining epistemic spot checks and prediction markets sound super fun to you? Good news: We're launching round three of the experiment today, with prizes of up to $65/question. The focal book will be The Unbound Prometheus, by David S. Landes, on the Industrial Revolution. The market opens today and will remain open until 10/27 (inclusive).

How to Improve Your Sleep

This piece is cross-posted on my blog here.

Do you feel sleepy enough during the day that it impairs your ability to function? Do you fall asleep involuntarily during the day? Do you feel tired enough that you feel weak, have difficulty doing things, or feel unable to focus?

If you answered yes to any of these questions, here's a set of guidelines to help you figure out what's worth trying. I go over generally applicable advice for daytime sleepiness and fatigue, then cover signs that would indicate more rare causes you can investigate. If feeling tired is significantly impairing your productivity, the information value of trying these experiments is probably well worthwhile.

This infographic checks for the most common causes of fatigue and what you should do it for each. These common causes account for the majority of people with daytime sleepiness, so there's a good chance they cover the most relevant information for you. If you've already read a bunch about sleep, you'll find loads more detail in the eleven pages of research that follow.

Lynette Bye presents

HOW TO MANAGE FATIGUE

A Simplified Decision Tree

Are you consistently giving yourself 7 to 9 hours in bed?

If not, start giving yourself enough time in bed. Try seeing how much time you need by sleeping without an alarm for several days.

If you are, try exercising regularly and improving your sleep hygiene. You should see an improvement within four weeks if you're consistently doing so.

Does it take you 30 minutes or more to fall asleep, or do you sleep six hours or less on three or more nights per week?

If so, check out self-administered therapies for insomnia. If those don't work within 4-6 weeks, talk to a doctor or therapist about CBT for insomnia and/or sleep aids.

Do you snore a lot or wake up gasping for breath (or does your partner report you doing either)?

If so, talk with a sleep specialist about sleep apnea.

Do you score mild or worse on the PHQ-9?

If so, talk to your doctor about the possibility you might have depression (even mild depression might make you fatigued).

If none of the above describe you

If you're excessively sleepy despite getting lots of sleep and none of the above address your situation, you should talk to a doctor. Check out the sections below on health issues and sleep disorders for more information.

_This guide was heavily influenced by UpToDate.com, an evidence-based clinical resource, and interviews with medical professionals. But, just so we're clear, I'm not a medical professional and these suggestions aren't medical advice. Listen to the person who actually went to medical school if they give you different advice. _

1. Are you fatigued?

If you can fall asleep during the day in less than 8 minutes, you're probably sleep-deprived and may have a sleep disorder. You might have fatigue (or excessive daytime sleepiness) if:

- You feel sleepy enough during the day that it impairs your ability to function, or you fall asleep involuntarily during the day.
- You feel tired enough that you feel weak, have difficulty doing things, or can't focus because you're tired.

If you experience the above more than once a week or have some weeks where you're just tired the whole time, it's probably worthwhile for you to spend time trying to solve the issue. (It's probably not fatigue if you just feel tired right after a trip when jet-lagged, or if you had to pull an all-nighter.)

What works to make you feel well-rested will probably be different from others, particularly people who always feel rested by default. So, you can approach your sleep as a hits-based experiment. It might take you some trial and error to find what works for you, but one big win justifies a lot of little experiments. So if sleep is a big problem, the information value of trying these experiments is quite high.

2. Are you giving yourself enough time to sleep?

First and foremost, are you giving yourself 7 to 9 hours in bed every night? If you're giving yourself less than 9 hours in bed, you're sleeping through the night, and you're tired during the day, then we found your problem - or at least the first problem you need to address.

Adults usually need between 7 and 9 hours of sleep per night, and regularly sleeping outside the normal range may be a sign of a serious health problem or, if you're doing it voluntarily, may compromise your health and performance. Spending more time in bed can be counterproductive if you have insomnia, but you probably don't have insomnia if you're sleeping soundly through the night.

If you aren't sure spending the extra time asleep is worth it, I encourage you to read Why We Sleep. For now, we'll content ourselves with this tongue-in-cheek summary of the new wonder drug, sleep, which "makes you live longer. It enhances your memory and makes you more creative. It makes you look more attractive. It keeps you slim and lowers food cravings. It protects you from cancer and dementia. It wards off colds and the flu. It lowers your risk of heart attacks and stroke, not to mention diabetes. You'll even feel happier, less depressed, and less anxious."

Start by seeing how long you sleep without an alarm when you can afford to let yourself sleep in. This experiment is good for checking how much sleep you need. In

general, waking up at a consistent time is a good sleep hygiene practice. It may take you several days of sleeping more than you normally need to work off the sleep debt before you can tell how much sleep you normally need. This could completely solve the problem, and it'll save you wasting time having a doctor tell you to try sleeping more.

3. Are you exercising regularly?

Exercise for at least two hours per week, ideally 30 minutes per day. Regular exercise has small-to-medium beneficial effects on total sleep time, sleep efficiency, sleep latency, and sleep quality, plus it can reduce fatigue in general. It can also be good for waking up midday if you're feeling sleepy. Both cardio and strength training will help, so just pick a form of exercise you enjoy.

It seems fine to exercise even if you'll go to bed in a couple of hours. While this meta-analysis found that evening exercise didn't have a significant impact on sleep latency, sleep efficiency, or total sleep time, it suggests that vigorous exercise within one hour of bed might keep you awake longer and make you sleep less. My best guess is that it's somewhat better to exercise earlier in the day, but better to do it close to bed than not at all if you can't do it earlier.

4. Are you practicing solid sleep hygiene?

According to Doctors Bertisch and Peña, you should see noticeable improvement within four to six weeks of consistent practice if sleep hygiene is going to help you. If you've already spent a lot of effort to improve your sleep hygiene and it didn't work, it might be worth skipping to the other sections.

The short summary of tips:

1. Wake up at the same time every day, including weekends.
2. Don't take stimulants closer to your bedtime than your body clears them out - a rough guess is about 10 hours before bed for 1 cup of coffee, and at least 15 hours before bed for 100mg of modafinil.
3. Don't drink alcohol close to your bedtime - probably good to stop drinking 4 or 5 hours before bed.
4. Avoid blue light for a couple of hours prior to your bedtime. This includes all LED lights and flat screens from digital devices. Turning off all LEDs and screens might be a challenge, but you can somewhat reduce the exposure on screens by using blue-light reducing programs such as Flux and using Warm White lights in the evening.
5. Expose yourself to bright light shortly after waking up, ideally natural, full-spectrum light.
6. Take 0.3mg of melatonin 30 minutes before you go to bed.
7. Sleep in a cool room, between 60-67 degrees Fahrenheit.
8. Take a warm shower before bed.
9. Make your bedroom as quiet as possible at night, particularly if you notice noise is waking you up.

The long version of these tips:

The feeling of sleepiness

Before I talk about standard sleep hygiene tips, some context on how sleep works. Two factors govern the experience of sleepiness: your circadian rhythm and homeostatic sleep drive. First, the circadian rhythm aligns your desire to sleep, be awake, and even when to eat, with your inner biological clock. Now, this clock checks if it's light out, and adjusts it's timing a bit to match the daylight. When it sees it's dark out, it releases melatonin to let your body know it's time to sleep now. Then in the morning when light filters in through your eyelids, melatonin gets shut off to wake you up. The circadian rhythm can only adjust itself a bit each day, which is why you feel jetlagged for a few days when you try to suddenly change your rhythm by several hours.

Meanwhile, your homeostatic sleep drive builds pressure to sleep by dialing down alert-promoting brain regions and dialing up sleep-inducing parts of the brain. A key part of this drive, the chemical adenosine, increases each minute you're awake. After 12 to 16 hours of being awake, adenosine will be enough to make you irresistibly sleepy. You can temporarily pause the experience of sleepiness caused by adenosine by consuming caffeine, but the full force of the built up adenosine will hit you when the caffeine wears off. So be ready to crash hard.

These two forces act independently. If you're sleeping a normal schedule, they will line up; your circadian rhythm will signal you to go to bed around when the adenosine is getting high enough to demand sleep, and the adenosine will be low after a good night's sleep when your circadian rhythm is signaling to rise and shine. If you pull an all-nighter or are jetlagged, then these two forces may act on you at different times.

Together, these forces trigger a host of other chemicals that move you along the spectrum of sleep and wakefulness. Along this spectrum, excessive daytime sleepiness is deficient daytime arousal, and insomnia is excessive nighttime arousal.

So, roughly in order from most to least likely to help:

Wake up at the same time

So, your circadian rhythm is pretty important. If you don't sleep at consistent times, you increase the risk that you're not sleeping at the times when your circadian rhythm thinks you should be and vice versa. A consistent sleep schedule can help these line up.

Ideally, you would go to bed at the same time each night, and wake up around the same time each morning. Sleeping on a consistent schedule is one of the more difficult sleep hygiene tips. Both Dr. Bertisch and stimulus control therapy emphasize waking up at the same time, so I'd focus on this one if you want to pick just one.

Also, don't use snooze on your alarm (if you have to use an alarm); fragmenting your sleep is worse than waking once.

Avoid stimulants and alcohol too close to bedtime

Too close to bedtime will differ by person, but avoiding caffeine after lunch is a good rule of thumb unless you've experimented and found different results for you personally.

For caffeine, one study found 400mg 6 hours before bed still had a huge impact on sleep - the caffeine reduced average total sleep time by an hour and reduced average sleep efficiency 9% compared to the placebo. It also increased the average sleep latency by 24 minutes (from 20 to 44) (though this result only had a p-value of .13).

That makes sense based on the half-life of caffeine. It takes about 5 hours to clear half the caffeine out of your system, so that means that after 6 hours there was still close to 200mg of caffeine. It's understandable that the equivalent of two cups of coffee would make the study participants feel more alert while they're trying to sleep. If we assume that 25mg of caffeine won't interfere with sleep, then you would need to leave at least 15 hours between taking 400mg and going to bed, and at least 10 hours for 100mg. For stimulants with longer half-lives, you would need to take it a correspondingly longer time before bed (e.g. for modafinil, half of it will still be in your system 13 hours or more after taking the drug).

It's probably worth experimenting to see how long you're impacted by stimulants - the time it takes your body to eliminate them can vary a lot. Also, according to a nurse I spoke with, your brain desensitizes to the overload of chemicals, so the acute alertness effect is probably gone after around 15 hours even if you still have a dose in your system. However, there may still be residual effects from the remaining dose, e.g. trouble sleeping.

Similarly, drinking alcohol shortly before bed has a net negative impact on sleep. Alcohol reduces sleep latency (time to fall asleep) but decreases sleep quality and delays REM sleep. High doses of alcohol (4+ standard drinks) had a significant negative impact on sleep even when taken 4 hours before bed, but drinking 5 hours before bed was significantly better than drinking just before bed.

Have a dark night and bright morning

Bright light, particularly blue light, delays melatonin production. Delaying melatonin production keeps you awake longer immediately, and can reset your circadian rhythm to a later time over a longer time scale. Electronic devices and LED bulbs emit a lot of blue light, so we usually get too much blue light late at night.

Turning off your electronics 2-3 hours before bed, using flux or other programs that reduce blue light from your devices at night, using Warm White bulbs, and wearing blue-light blocking goggles might reduce the disturbance. Taking a melatonin supplement before bed might help correct the delayed melatonin production with far less hassle, but the sleep specialists I asked were less encouraging. So probably worth at least testing flux in addition to melatonin.

Darkness while sleeping might reinforce that it's time to sleep. In order from least to most hassle: Buy a sleep mask. Install blackout blinds. Cover all of the little lights in your room with black tape (e.g. the smoke detector light). If you don't have blackout blinds, try moving your bedtime earlier to see if you sleep more when the sun isn't waking you up (it takes a week or two to adjust your bedtime). This last idea didn't work for me, but the sleep specialist thought it might help when trying to sleep more, perhaps because the extra sleep time is at night instead of during daylight.

Bright light in the morning will probably increase your alertness by reducing melatonin (small to medium effect size). If your room doesn't have bright natural sunlight in the morning, lots of bright bulbs can simulate the effect. I have 10,000 Lumens in my light fixture (I affectionately call it the sun), and I know people with way more.

Take Melatonin

Melatonin isn't a standard sleep tip but seems like a mostly safe experiment to try.

According to Slate Star Codex and Gwern, taking melatonin before bed will probably help you fall asleep more quickly and improve your sleep quality. The studies Gwern quotes found melatonin reduced the average sleep latency (the time it takes to fall asleep) by 3.9 to 16.8 minutes, a comparable amount to sleep pills. Furthermore, sleep efficiency increased by 3.1%, and sleep duration increased by 13.7 min.

However, the sleep specialists were more pessimistic that melatonin would help as a sleep aid. Dr. Cooper thought taking melatonin was fine if you also shut off blue screens. Dr. Bertisch thought melatonin probably wouldn't help with insomnia, but it was "not unreasonable" to try. Dr. Zhou wrote that the "evidence for melatonin as a hypnotic is weak" and "there have no been long-term studies of the safety of melatonin, and there are studies showing that the non-regulated OTC industry often does not sell what it promises to when it comes to melatonin."

I looked into the worries that melatonin isn't regulated by the FDA, and that commercial supplements of melatonin have been found to contain more or less than the amount claimed on the label. Both are true, but it's probably not a big deal. The study found that >80% of the supplements studied differed from the amount on the label by less than 50%. If you're aiming for 0.3 MG, you're going to be fine if it differs from that amount by 50%. If you notice that you sleep worse after taking even a small dose, you might want to take even less in case you're accidentally getting too much. However, there aren't long term studies, so use your judgment about taking melatonin regularly over periods longer than three months.

For use as a simple hypnotic (sleep aid), melatonin is most commonly taken shortly before bed to immediately help you sleep (Gwern recommends taking melatonin 30min before you want to sleep). The recommended dose is around 0.3mg or less. However most pills contain at least 10 times that amount, so check the amount before you purchase for one with 300MCG like this one.

Sleep cool

The core body temperature naturally drops at night as skin blood flow increases, drawing body heat away from the core.

This has been experimentally tested by slightly warming skin so that more blood flowed there, causing the core temperature to drop and the participants to fall asleep more quickly. One study found that slightly warming the skin (0.4°C) reduced the sleep latency (time to fall asleep) by 1.84 minutes in healthy participants and 2.85 minutes in insomniacs. Taking a warm shower or splashing warm water on your face before bed may have the same impact.

A cool bedroom is presumed to improve sleep by mimicking this temperature drop. Most of the sources I looked at suggested between 60-67 degrees Fahrenheit, apparently because some studies indicated hot rooms decreased sleep quality. Open the window or turn on an AC if it's too warm. I've heard a few strong recommendations for a ChiliPad to maintain the proper sleep temperature.

Keep the noise down

Try keeping your sleeping area quiet to avoid decreasing sleep quality. This study found that traffic noises between 39 and 50 dBA (with a maximum level up to 74 dBA) reduced total sleep time by 16 minutes and decreased sleep quality compared to a 32 dBA pink noise control. 32 dBA is about a whisper, while the other sounds range from average home noise to inside a car going 60 miles per hour. This larger study found that there were no significant changes in sleep structure if the sleep disturbances "did not exceed 4×80 dB(A), 8×70 dB(A), 16×60 dB(A), 32×55 dB(A) and 64×45 dB(A) in a single night." So a good target seems to be noise levels near whisper levels while you sleep (or at least under normal conversation levels), with few spikes in noise levels.

Try earplugs or white noise if noise bothers you. You can test white noise cheaply on your phone, and buy a white noise machine if you like it. If you want to go a step further, install soundproofing. Insulating around doors and windows (anywhere air can get through) seems particularly important.

5. Are you depressed?

Don't skip this right away, even if you think the answer is probably no. Depression affects 1 in 15 people, many of whom are not aware that they are experiencing depression. (I've known several people who didn't realize they were depressed, sometimes for many years.) So I encourage you to err on the side of taking the PHQ-9, a short screening tool.

Only a doctor can diagnose you with depression, but if you're feeling miserable all the time or you score high on the scale, you can probably make a pretty good guess. You might want to retake the scale every week for a few weeks if you're on the low end – even mild depression some of the time may be worth talking with a doctor about.

If you're concerned, go see a doctor. Your doctor can prescribe medications that might help a lot. Wellbutrin (generic bupropion) might be a good one to try if fatigue is one of your main symptoms.

If you think you might have depression but aren't ready to see a doctor, I recommend this Slate Star Codex article.

6. Are you having trouble falling or staying asleep? Insomnia

If you give yourself plenty of time in bed but you can't fall asleep or you wake up during the night, you probably have insomnia. It might often take you 30 minutes or more to fall asleep or you might sleep six hours or less on three or more nights per week despite spending enough hours in bed. You probably also won't be able to nap during the day, despite being tired.

If that's you, then the best place to start is sleep therapy and sleep hygiene. Behavioral therapies can help with falling asleep and managing waking up in the night. CBT for Insomnia (CBT-I) can include relaxation therapy, stimulus control therapy, cognitive therapy, phototherapy for adjusting sleep times, and sleep restriction.

You can experiment with these yourself at home, but talk to a doctor if you aren't seeing results after four to six weeks. According to Dr. Bertisch, "Self-help books for CBT-I are helpful, so it's something many people can do on their own." However, she warns that patients could become discouraged if the self-therapy isn't working and be less likely to talk to a doctor, despite the fact that a doctor can often help even when self-help failed.

At a high-level overview, therapy reduces insomnia the most, followed by sleep drugs, followed by sleep hygiene. However, it seems likely that there is a selection effect in the insomnia cases that get included in sleep studies. Dr. Bertisch said that most people who got to the point of seeing a sleep doctor were already practicing good sleep hygiene, with the most common exceptions being not having a consistent sleep schedule and using blue-light-emitting electronic devices at night. So starting with both self-therapy and sleep hygiene seems reasonable. Again, you should see results within four to six weeks of consistent use if sleep hygiene and self-therapy are going to work.

CBT-I

CBT-I is considered the gold standard for treating insomnia. Studies find it to be more effective than sleep medication for reducing insomnia. It's delivered one on one with a therapist, so it may be bottlenecked on finding a therapist who specializes in CBT-I. However, you can try the individual therapies yourself.

One meta-analysis found that CBT for insomnia reduced sleep latency by 19 minutes, increased total sleep time by 7 minutes, and increased sleep efficiency by 9% compared to the placebo.

Stimulus Control Therapy

This one is probably a safe option to try first, either alone or in combination with others. It aims to form an association between being in bed and sleeping (rather than being awake!), and set a consistent sleep schedule. The guidelines are as follows:

1. Lie down to go to sleep only when you are sleepy.
2. Do not use your bed for anything except sleep; that is, do not read, watch television, eat, or worry in bed. Sexual activity is the only exception to this rule. On such occasions, the instructions are to be followed afterward when you intend to go to sleep.
3. If you find yourself unable to fall asleep, get up and go into another room. Stay up as long as you wish and then return to the bedroom to sleep. Although we do not want you to watch the clock, we want you to get out of bed if you do not fall asleep immediately. Remember, the goal is to associate your bed with falling asleep quickly! If you are in bed for more than about 10 minutes without falling asleep and have not gotten up, you are not following this instruction.
4. If you still cannot fall asleep, repeat step (3). Do this as often as is necessary throughout the night.
5. Set your alarm and get up at the same time every morning irrespective of how much sleep you got during the night. This will help your body acquire a consistent sleep rhythm.
6. Do not nap during the day.

Phototherapy

Phototherapy uses light to shift your sleep cycle if you have a circadian rhythm disorder such as sleeping and waking up too late or too early. You can read more about shifting your sleep schedule on Slate Star Codex's melatonin article.

Sleep restriction therapy

You limit total time in bed so that you're tired when you go to bed and will sleep well. It usually starts with limiting you to around 6 hours in bed per night, and then gradually increases as long as your sleep efficiency stays above 85%. It can work well, but if done improperly, this therapy can leave you sleep deprived.

Cognitive therapy

Cognitive therapy with a therapist can work to reduce anxiety, including anxiety about not being able to sleep which makes it harder to sleep. I can imagine that it might be worth trying other methods of releasing stress and worry before bed, such as journaling or meditation, if anxiety is a common cause of insomnia for you.

Progressive relaxation therapy

You progressively tense and relax each muscle group over about 45 minutes. It seems to be considered less effective but is easy to try on your own.

Sleep aids

If therapy and sleep hygiene isn't working, you can try hypnotics (sleep drugs). They have a lot of side effects but are occasionally the only thing that works. If you are going to use a sleep aid, it's best to stick with the lowest dose for the shortest time that you can manage.

According to a pharmacology book, there are new drugs that avoid most of the problems of the old drugs, which included causing daytime sleepiness or wake maintenance problems, loss of efficacy over time, and causing worse insomnia when ceasing the drug. However, hypnotics may not be much better than Melatonin, and they can have weird side effects like causing the user to eat, drive, or have sex while asleep on the drugs with no memory of it the next day.

All that said, you can talk to your doctor if you want to try them, but it shouldn't be your first - or third - line of action.

7. Do you have a health issue causing fatigue?

You'll need to see your doctor if you have health concerns. If you suddenly started feeling fatigued within the past three months, you should probably talk to your doctor. It may be a sign of a health issue, e.g. mononucleosis and depression are both characterized by a relatively sudden onset of fatigue.

Your doctor will probably do blood tests to check for two common causes of fatigue, anemia and hypothyroidism, and possibly vitamin deficiencies. If your doctor doesn't do blood tests, you can ask them politely about these. Your doctor can also check your medications to see if one of them, or a combination of them, is making you excessively sleepy.

If your fatigue is caused by other health issues, things get a bit trickier. Your doctor will probably be able to help you decide which things to test for based on your other symptoms. E.g. if you also faint a couple of times a day, your fatigue might be caused by POTS.

Finally, your primary care doctor can make other recommendations as appropriate, including if you should see a specialist or try sleep aids.

8. Do you have a sleep disorder?

If you have a sleep disorder, then the other things are unlikely to fix the problem, and you'll want to see a sleep specialist to get it diagnosed. Note: You may have a sleep disorder and not know it. One of the sleep specialists I emailed, Dr. Zhou, said he sees "more than enough people in clinic with sleep apnea who think that shutting off electronics will help them."

Sleep apnea is the most common sleep disorder - estimates of Americans afflicted range from 3% to 7%. If you feel fatigued and snore a lot or experience waking up gasping for breath (or if your partner reports you doing either), then you should get checked for sleep apnea. While not everyone who snores has sleep apnea, almost everyone who has sleep apnea snores.

Similarly, if you score moderate or severe on the Epworth Sleepiness Scale, you have a higher chance of having narcolepsy, although it is rare. If you have an involuntary urge to move your legs while lying still, you may have restless leg syndrome. If you can't fall asleep until too late at night or always fall asleep too early, you might have a circadian rhythm disorder.

If you have sleep apnea, narcolepsy, or a handful of other specific sleep disorders, a sleep specialist will be quite helpful. A sleep specialist may order a Polysomnogram and Multiple Sleep Latency Test, where they monitor your sleep overnight and then during 5 naps the next day to check for disorders such as sleep apnea and narcolepsy. If they are pretty sure you have moderate to severe sleep apnea, they can do the test at home instead of in a lab.

I've spoken to a handful of people (4-5) who didn't have symptoms of specific disorders and spoke to a sleep specialist, and none of them found it helpful. UpToDate says that with many patients the cause of excessive sleepiness is apparent, but others should do formal testing, particularly if they show signs of sleep disorders. So basically, you might have a sleep disorder and not know it, but there's a good chance you and your primary care doctor can predict in advance whether it will be valuable to visit a sleep specialist.

9. What else can you do during the day to manage energy?

Identifying the cause of your fatigue is valuable for managing it long term. In the meantime, however, there are things you can do during the day to manage your energy.

Stimulants

Stimulants can be helpful, particularly for short periods or cyclically. Some build up tolerance or dependence quickly, such as caffeine. So taking them when needed or cyclically might be better than every day. Again, it's probably best to take stimulants earlier in the day. You probably also want to do your sleep experiments before you switch to stimulants, so you can more easily tell if they are working.

For caffeine, L-Theanine is supposed to balance the jitteriness, generally a 2:1 ratio of L-Theanine to caffeine. A lot of caffeine pills come as a caffeine-L-Theanine combination. You can also buy L-Theanine separately if you drink your caffeine. Caffeine basically tricks you into not feeling sleepy by blocking adenosine, a molecule that induces the feeling of sleepiness. This can lead you to crash when the caffeine wears off because you suddenly experience all the adenosine that built up.

You could talk to your doctor about a modafinil prescription if your fatigue is extreme. Although modafinil is useful for a variety of cognitive enhancements in non-sleep deprived populations, it is unlikely to be prescribed for that purpose. (You can get a large discount via GoodRX if your insurance doesn't cover it.) Gwern has a detailed post about modafinil here.

This Slate Star Codex post has more information on potential stimulants to try.

Short naps

Short naps might help, but be careful of longer naps further confusing your biological clock. Naps can alleviate fatigue when you are sleep deprived, but they can also increase insomnia if you already struggle with it.

A 15-minute nap is probably a good baseline to start with (it has been shown to work well, and is a good length to let you rest without entering deep sleep), and then you can experiment to see what is the ideal time for you. Some recommendations suggest trying taking caffeine with the nap, so the caffeine will be kicking in as you wake up 15 minutes later. Again, I suggest being careful not taking stimulants less than two half-lives-ish before you want to sleep.

Sources and Notes

You can check out this supplemental document for my sources and notes if you want even more information.

If you want more information, I recommend UpToDate.com for detailed guides on fatigue, insomnia, and relevant treatments. Why We Sleep also contains a plethora of information about sleep and why it's so important.

Special thanks to the medical professionals who made time to talk or email with me regarding this article, especially Dr. Suzanne Bertisch, Clinical Director of Behavioral Sleep Medicine at Brigham and Women's Hospital and Assistant Professor of Medicine

at Harvard Medical School; Dr. Laura Barger, Instructor in Medicine at the Harvard Medical School Devision of Sleep Medicine; Dr. Victor Peña, Surgeon and Health Coach; Lucas Valenca, RN; Dr. Joanna Cooper, Neurologist and Sleep Specialist; Dr. Eric Zhou, Psychologist and Faculty in the Division of Sleep Medicine at Harvard Medical School.

What empirical work has been done that bears on the 'freebit picture' of free will?

This picture was described in Scott Aaronson's essay The Ghost in the Quantum Turing Machine in 2013, and claims that human free will is related to our choices being caused by (and/or causing) quantum bits from the initial state of the universe that first have macroscopic effects inside our brains, where all other observers must have purely Knightian uncertainty over such bits[*]. For it to be plausible, a few facts have to be true about human brain biology and cosmic background radiation:

1. "Quantum uncertainty—for example, in the opening and closing of sodium-ion channels—can not only get chaotically amplified by brain activity, but can do so “surgically” and on “reasonable” timescales."
2. All photons that impinge on human brains have quantum states that could not "be altered, maintaining a spacetime history consistent with the laws of physics, without also altering classical degrees of freedom in the photons’ causal past".

Have these questions been studied in the intervening years, and what have the results been? Note that the plausibility of the picture has been discussed before on LW, and *I'm not interested in further discussing whether a priori it seems at all promising to link free will and Knightian uncertainty*.

[*] This is a poor summary, I recommend reading the paper if you have time.

AI Alignment Open Thread October 2019

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

Continuing the experiment from August, let's try another open thread for AI Alignment discussion. The goal is to be a place where researchers and upcoming research can ask small questions they are confused about, share early stage ideas and have lower-key discussions.

The sentence structure of mathematics

"Alice pushes Bob."

"Cat drinks milk."

"Comment hurts feelings."

These are all different sentences that describe wildly different things. People are very different from cats, and cats are very different from comments. Bob, milk, and feelings don't have much to do with each other. Pushing, drinking, and (emotionally) hurting are also really different things.

But I bet these sentences all feel really similar to you.

They should feel similar. They all have the same *structure*. Specifically, that structure is

Noun verb noun.

Because these sentences all share the same fundamental underlying structure, they all feel quite similar even though they are very different on the surface. (The mathematical term for "fundamentally the same but different on the surface" is *isomorphic*.)

When you studied sentence structure back in grammar school (it wasn't just me, right?) you learned to break down sentences into their *parts of speech*. You learn that nouns are persons, places, or things, and verbs are the activities that nouns do. Adjectives describe nouns, and adverbs describe pretty much anything. Prepositions tell you where nouns go. Etc.

Parts of speech are really *abstract* and really *general*. When you look at the surface, the sentence

the ant crawls on the ground

and the sentence

the spaceship flies through space

could not possibly be more different. But when you look at the sentence structure, they're nearly identical.

The concept of "parts of speech" emerge when we notice certain general patterns arising in the way we speak. We notice that whether we're talking about ants or spaceships, we're always talking about *things*. And whether we're talking about crawling or flying, we're always talking about *actions*.

And so on for adjectives, adverbs, conjunctions, etc., which always seem to relate back to nouns and verbs—adjectives modify nouns, for example.

Next we simply give things and actions, descriptors and relational terms some confusing names to make sure the peons can't catch on—nouns and verbs, adjectives and prepositions—and we have a way of breaking down *any English sentence* into its *fundamental parts*.

That is to say, if you know the abstract rules governing sentence structure—the types of pieces and their connections—you can come up with structures that *any English sentence* is but a *particular example of*.

Like how "Alice pushes Bob" is but a particular example of "Noun verb noun."

At the most basic level, category theory breaks down mathematics into its parts of speech. It turns out that mathematics is pretty much just nouns and verbs at its simplest—just like how, if you read between the lines a bit, any English sentence can be boiled down to its nouns and verbs. Those are the "main players" which everything else just modifies in some fashion.

In mathematics, a noun is called an *object*.

A verb is called a *morphism* or *arrow*. We'll explore the terminology of morphism a bit more next time. As to why they can also be called arrows, that's because verbs appear to have directions: One noun does the verb, and another noun (potentially the same noun, like pinching yourself) receives the verb. So you could draw that as an arrow like so:

$$\text{Alice} \xrightarrow{\text{push}} \text{Bob}.$$

This is exactly how we diagram objects and morphisms in category theory, with one difference: we typically use single letters in place of full names. (I'd explain the value of concision here, but it seems hypocritical.) So if Alice and Bob are objects in our category, and Alice's push of Bob is the morphism, then we might write it this way:

$$A \xrightarrow{p} B.$$

Equally legitimate is to highlight the morphism up front. (We'll see they're the real stars of the show):

$$p : A \to B.$$

So now you understand objects and morphisms, the basic pieces of any category, just like how nouns and verbs are the basic pieces of any sentence.

Of course, making a sentence isn't as simple as mashing nouns and verbs together. We need to make sure that the sentence makes sense. To paraphrase Harrison Ford, you can *write* "colorless green ideas sleep furiously", but you sure can't think it.

We'll explore the rules that define a category in the next post.

[AN #70]: Agents that help humans who are still learning about their own preferences

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Highlights

The Assistive Multi-Armed Bandit *(Lawrence Chan et al)* (summarized by Asya): Standard approaches for inverse reinforcement learning assume that humans are acting optimally according to their preferences, rather than learning about their preferences as time goes on. This paper tries to model the latter by introducing the *assistive multi-armed bandit* problem.

In the standard *multi-armed bandit* problem, a player repeatedly chooses one of several “arms” to pull, where each arm provides reward according to some unknown distribution. Imagine getting 1000 free plays on your choice of 10 different, unknown slot machines. This is a hard problem since the player must trade off between exploration (learning about some arm) and exploitation (pulling the best arm so far). In *assistive multi-armed bandit*, a robot is given the opportunity to intercept the player every round and pull an arm of its choice. If it does not intercept, it can see the arm pulled by the player but not the reward the player receives. This formalizes the notion of an AI with only partial information trying to help a learning agent optimize their reward.

The paper does some theoretical analysis of this problem as well as an experimental set-up involving a neural network and players acting according to a variety of different policies. It makes several observations about the problem:

- A player better at learning does not necessarily lead to the player-robot team performing better-- the robot can help a suboptimal player do better in accordance with how much information the player's arm pulls convey about the reward of the arm.

- A robot is best at assisting when it has the right model for how the player is learning.

- A robot that models the player as learning generally does better than a robot that does not, even if the robot has the wrong model for the player's learning.

- The problem is very sensitive to which learning model the player uses and which learning model the robot assumes. Some player learning models can only be effectively assisted when they are correctly modeled. Some robot-assumed learning models effectively assist for a variety of actual player learning models.

Asya's opinion: The standard inverse reinforcement learning assumption about humans acting optimally seems unrealistic; I think this paper provides an insightful initial step in not having that assumption and models the non-optimal version of the problem in a clean and compelling way. I think it's a noteworthy observation that this problem is very sensitive to the player's learning model, and I agree with the paper that this suggests that we should put effort into researching actual human learning strategies. I am unsure how to think about the insights here generalizing to other inverse reinforcement learning cases.

Technical AI alignment

Problems

Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence *(David Manheim)* (summarized by Flo): While Categorizing Variants of Goodhart's Law explains failure modes that occur when a single agent's proxy becomes decoupled from the true goal, this paper aims to characterize failures involving multiple agents:

Accidental steering happens when the combined actions of multiple agents facilitate single-agent failures. For example, catching more fish now is usually positively correlated with a fisherman's long term goals, but this relationship inverts once there are lots of fishermen optimizing for short term gains and the fish population collapses.

Coordination Failure occurs when agents with mutually compatible goals fail to coordinate. For example, due to incomplete models of other agent's goals and capabilities, two agents sharing a goal might compete for a resource even though one of them is strictly better at converting the resource into progress towards their goal.

Adversarial optimization is when an agent **O** steers the world into states where **V**'s proxy goal is positively correlated with **O**'s goal. For example, one could exploit investors who use short term volatility as a proxy for risk by selling them instruments that are not very volatile but still risky.

Input Spoofing is the act of one agent manipulating another learning agent's model, either by manufacturing false evidence or by filtering the received evidence systematically, as arguably happened with Microsoft's Tay.

Finally, **Goal co-option** happens when agent **O** has (partial) control over the hardware agent **V** runs or relies on. This way, **O** can either modify the reward signal **V** receives to change what **V** optimizes for, or it can directly change **V**'s outputs.

The difficulties in precisely modelling other sophisticated agents and other concerns related to embedded agency make it hard to completely avoid these failure modes with current methods. Slowing down the deployment of AI systems and focussing on the mitigation of the discussed failure modes might prevent limited near term catastrophes, which in turn might cause a slowdown of further deployment and prioritization of safety.

Flo's opinion: I like that this paper subdivides failure modes that can happen in multiparty optimization into several clear categories and provides various models and examples for each of them. I am unsure about the conclusion: on one hand, slowing down deployment to improve the safety of contemporary systems seems very

sensible. On the other hand, it seems like there would be some failures of limited scope that are hard to reproduce "in the lab". Widely deployed AI systems might provide us with valuable empirical data about these failures and improve our understanding of the failure modes in general. I guess ideally there would be differential deployment with rapid deployment in noncritical areas like managing local parking lots, but very slow deployment for critical infrastructure.

Rohin's opinion: I'm particularly interested in an analysis of how these kinds of failures affect existential risk. I'm not sure if David believes they are relevant for x-risk, but even if so the arguments aren't presented in this paper.

Mesa optimization

Relaxed adversarial training for inner alignment *(Evan Hubinger)* (summarized by Matthew): Previously, Paul Christiano proposed creating an adversary to search for inputs that would make a powerful model behave "unacceptably" and then penalizing the model accordingly. To make the adversary's job easier, Paul relaxed the problem so that it only needed to find a pseudo-input, which can be thought of as predicate that constrains possible inputs. This post expands on Paul's proposal by first defining a formal unacceptability penalty and then analyzing a number of scenarios in light of this framework. The penalty relies on the idea of an amplified model inspecting the unamplified version of itself. For this procedure to work, amplified overseers must be able to correctly deduce whether potential inputs will yield unacceptable behavior in their unamplified selves, which seems plausible since it should know everything the unamplified version does. The post concludes by arguing that progress in model transparency is key to these acceptability guarantees. In particular, Evan emphasizes the need to decompose models into the parts involved in their internal optimization processes, such as their world models, optimization procedures, and objectives.

Matthew's opinion: I agree that transparency is an important condition for the adversary, since it would be hard to search for catastrophe-inducing inputs without details of how the model operated. I'm less certain that this particular decomposition of machine learning models is necessary. More generally, I am excited to see how adversarial training can help with inner alignment.

Learning human intent

Learning from Observations Using a Single Video Demonstration and Human Feedback *(Sunil Gandhi et al)* (summarized by Zach): Designing rewards can be a long and consuming process, even for experts. One common method to circumvent this problem is through demonstration. However, it might be difficult to record demonstrations in a standard representation, such as joint positions. **In this paper, the authors propose using human feedback to circumvent the discrepancy between how demonstrations are recorded (video) and the desired standard representation (joint positions).** First, humans provide similarity evaluations of short clips of an expert demonstration to the agent's attempt and a similarity function is learned by the agent. Second, this similarity function is used to help train a policy that can imitate the expert. Both functions are learned jointly. The algorithm can learn to make a Hopper agent back-flip both from a Hopper demonstration of a back-flip, and from a YouTube video of a human backflipping. Ultimately, the authors show that their method improves over another method that uses human feedback without direct comparison to desired behavior.

Zach's opinion: This paper seems like a natural extension of prior work. The imitation learning problem from observation is well-known and difficult. Introducing human feedback with a structured state space definitely seems like a viable way to get around a lot of the known difficulties with other methods such as a GAIL.

Handling groups of agents

Collaborating with Humans Requires Understanding Them *(Micah Carroll et al)* (summarized by Rohin): *Note: I am second author on this paper.* Self-play agents (like those used to play Dota (AN #13) and Starcraft (AN #43)) are very good at coordinating with *themselves*, but not with other agents. They "expect" their partners to be similar to them; they are unable to predict what human partners would do. In competitive games, this is fine: if the human deviates from optimal play, even if you don't predict it you will still beat them. (Another way of saying this: the minimax theorem guarantees a minimum reward *regardless* of the opponent.) However, in cooperative settings, things are not so nice: a failure to anticipate your partner's plan can lead to arbitrarily bad outcomes. We demonstrate this with a simple environment that requires strong coordination based on the popular game Overcooked. We show that agents specifically trained to play alongside humans perform much better than self-play or population-based training when paired with humans, both in simulation and with a real user study.

Rohin's opinion: I wrote a short blog post talking about the implications of the work. Briefly, there are three potential impacts. First, it seems generically useful to understand how to coordinate with an unknown agent. Second, it is specifically useful for scaling up assistance games (AN #69), which are intractable to solve optimally. Finally, it can lead to more ML researchers focusing on solving problems with real humans, which may lead to us finding and solving other problems that will need to be solved in order to build aligned AI systems.

Read more: Paper: On the Utility of Learning about Humans for Human-AI Coordination

Learning Existing Social Conventions via Observationally Augmented Self-Play *(Adam Lerer and Alexander Peysakhovich)* (summarized by Rohin): This paper starts from the same key insight about self-play not working when it needs to generalize to out-of-distribution agents, but then does something different. They assume that the test-time agents are playing an **equilibrium policy**, that is, each agent plays a best response policy assuming all the other policies are fixed. They train their agent using a combination of imitation learning and self-play: the self-play gets them to learn an equilibrium behavior, while the imitation learning pushes them towards the equilibrium that the test-time agents use. They outperform both vanilla self-play and vanilla imitation learning.

Rohin's opinion: Humans don't play equilibrium policies, since they are often suboptimal. For example, in Overcooked, any equilibrium policy will zip around the layout, rarely waiting, which humans are not capable of doing. However, when you have a very limited dataset of human behavior, the bias provided by the assumption of an equilibrium policy probably does help the agent generalize better than a vanilla imitation learning model, and so this technique might do better when there is not much data.

Adversarial examples

Adversarial Policies: Attacking Deep Reinforcement Learning *(Adam Gleave et al)* (summarized by Sudhanshu): This work demonstrates the existence of *adversarial policies* of behaviour in high-dimensional, two-player zero-sum games. Specifically, they show that adversarially-trained agents ("Adv"), who can only affect a victim's observations of their (Adv's) states, can act in ways that confuse the victim into behaving suboptimally.

An adversarial policy is trained by reinforcement learning in a single-player paradigm where the victim is a black-box fixed policy that was previously trained via self-play to be robust to adversarial attacks. As a result, the adversarial policies learn to push the observations of the victim outside the training distribution, causing the victim to behave poorly. The adversarial policies do not actually behave intelligently, such as blocking or tackling the victim, but instead do unusual things like spasming in a manner that appears random to humans, curling into a ball or kneeling.

Further experiments showed that if the victim's observations of the adversary were removed, then the adversary was unable to learn such an adversarial policy. In addition, the victim's network activations were very different when playing against an adversarial policy relative to playing against a random or lifeless opponent. By comparing two similar games where the key difference was the number of adversary dimensions being observed, they showed that such policies were easier to learn in higher-dimensional games.

Sudhanshu's opinion: This work points to an important question about optimisation in high dimension continuous spaces: without guarantees on achieving solution optimality, how do we design performant systems that are robust to (irrelevant) off-distribution observations? By generating demonstrations that current methods are insufficient, it can inspire future work across areas like active learning, continual learning, fall-back policies, and exploration.

I had a tiny nit-pick: while the discussion is excellent, the paper doesn't cover whether this phenomenon has been observed before with discrete observation/action spaces, and why/why not, which I feel would be an important aspect to draw out. In a finite environment, the victim policy might have actually covered every possible situation, and thus be robust to such attacks; for continuous spaces, it is not clear to me whether we can *always* find an adversarial attack.

In separate correspondence, author Adam Gleave notes that he considers these to be relatively low-dimensional -- even MNIST has way more dimensions -- so when comparing to regular adversarial examples work, it seems like multi-agent RL is harder to make robust than supervised learning.

Read more: Adversarial Policies website

Other progress in AI

Reinforcement learning

Solving Rubik’s Cube with a Robot Hand *(OpenAI)* (summarized by Asya): Historically, researchers have had limited success making general purpose robot hands. Now, OpenAI has successfully trained a pair of neural networks to solve a Rubik's cube with a human-like robot hand (the learned portion of the problem is manipulating the hand

-- solving the Rubik's cube is specified via a classical algorithm). The hand is able to solve the Rubik's cube even under a variety of perturbations, including having some of its fingers tied together, or having its view of the cube partially occluded. The primary innovation presented is a new method called *Automatic Domain Randomization* (ADR). ADR automatically generates progressively more difficult environments to train on in simulation that are diverse enough to capture the physics of the real world. ADR performs better than existing domain randomization methods, which require manually specifying randomization ranges. The post speculates that ADR is actually leading to *emergent meta-learning*, where the network learns a learning algorithm that allows itself to rapidly adapt its behavior to its environment.

Asya's opinion: My impression is that this is a very impressive robotics result, largely because the problem of transferring training in simulation to real life ("sim2real") is extremly difficult. I also think it's quite novel if as the authors hypothesize, the system is exhibiting emergent meta-learning. It's worth noting that the hand is still not quite at human-level -- in the hardest configurations, it only succeeds 20% of the time, and for most experiments, the hand gets some of the state of the cube via Bluetooth sensors inside the cube, not just via vision.

Read more: Vox: Watch this robot solve a Rubik’s Cube one-handed

News

FHI DPhil Scholarships (summarized by Rohin): The Future of Humanity Institute will be awarding up to two DPhil scholarships for the 2020/21 academic year, open to students beginning a DPhil at the University of Oxford whose research aims to answer crucial questions for improving the long-term prospects of humanity. Applications will open around January or February, and decisions will be made in April.

Post-Doctoral Fellowship on Ethically Aligned Artificial Intelligence (summarized by Rohin) (H/T Daniel Dewey): Mila is looking for a postdoctoral fellow starting in Fall 2020 who would work on ethically aligned learning machines, towards building machines which can achieve specific goals while acting in a way consistent with human values and social norms. Applications are already being processed, and will continue to be processed until the position is filled.

Reflections on Premium Poker Tools: Part 3 - What I've learned

Previous posts:

- Part 1 - My journey
- Part 2 - Deciding to call it quits

I finally made the decision to call it quits. Now I think it would be a good time to reflect on my experiences and see if I could learn something from them.

Market size

As I talked about in the previous post, initially, I thought that the market size was hundreds of thousands of users, and that I could make something like $100/user. After talking to people in the industry, I now believe that the market is more like 5-10k users, and not all of them are willing to pay $100.

This is incredibly important! Going after a $20M market is very different from going after a $200k one. If I knew it was the latter in the beginning, I wouldn't have pursued this as a business. Spending 2+ years for the chance at making maybe $200k just isn't worth it, given the inherent uncertainty with startups, and the alternatives of pursuing a startup with a higher upside, or getting a job that pays approximately that much but does so with 100% certainty.

So what happened? Where did I go wrong? Let's see. This was (roughly) my initial logic:

> 100k subscribers to the poker subreddit. Educational YouTube videos gets 100k+ views. Popular posts on TwoPlusTwo (big poker forum) have 1M+ views over the years.
>
> The kind of person to subscribe to the poker subreddit, watch an educational YouTube video, or spend their time on TwoPlusTwo is probably somewhat serious about poker. They're trying to get better at the game. All of this is indicative of a market size of hundreds of thousands of users. Possibly more. And poker is expanding. So there seems to be a big market here.
>
> And poker players are probably pretty willing to spend money. Investing in software is +ROI. Poker players love ROI! And they tend to be on the wealthy side.

Maybe I really underestimated the divide between passive + free and active + paid. Watching a YouTube video is something that is passive. You sit there and consume information. Using poker software is something that is active. You have to sit and think and mess around with numbers. Watching a YouTube video is free. Poker software costs money.

But then what about the existence of poker books? There around 500 poker books on the market. The top ones get up to 100k sales, and many others get in the tens of

thousands, I think. Maybe it's that with books, someone is telling you what the right answers are, but with software, you have to figure out the right answers yourself.

Anyway, I think the bigger point is that I should have found people in the industry and asked them about the market size. I started doing that towards the end of my journey, but I should have done so from the beginning. People in the poker world all seem to have a pretty solid idea of what the market is really like. Why screw around trying to figure it out myself with these questionable proxies when I could just ask the people who actually know?

I really can't emphasize enough how huge this is. It would have only taken a few hours, and it would have saved me so much time.

So why didn't I go out and talk to people in the industry? I'm not quite sure. I think a part of it was that I didn't actually feel like I had to. It seemed pretty clear to me that the market was big, so I was more concerned with making the product awesome.

Another part of it is that I didn't see it as an option to talk to people in the industry. Because why would they want to talk to little ol' me? They're basically B-list celebrities, in some sense. They've written books. Tons and tons of people know them. Don't these people have hundreds of fan emails every day that they never respond to? That was my thinking in the beginning. Now, I've come to realize that they're just people, and that they're often happy to chat and provide advice.

It also would have been good if I had access to the actual financial data of my competitors. But none of them are public companies. Does that mean this isn't an option, or are there still ways?

I came across one cool approach last night. If you don't have access to their financial data, you could look at how many employees they have, and multiply by something like $125k or $200k. Something in that ballpark should give you an idea of their revenue. My competitors, from what I can tell, are all working solo. So that is a sign that they aren't making millions and millions of dollars. Not definitive, but definitely points in that direction.

Another interesting option is to actually call your competitors and talk to them! Eg. you could pretend to be a prospective employee, and in that conversation start asking about revenue and stuff. I'm not sure how I feel about that sort of deception though.

Here's a closing thought for this section: the world isn't that big. For the majority of my journey, my thoughts on the market size for poker has been, "I'm not sure exactly how big it is, but it's big enough." The world is just such a big place, and so many people play poker. The market just *has* to be huge. Now I realize that this isn't true.

"Just another month or two"

As I explain in the first post, for a very long time, I kept thinking to myself:

> I'm not sure if I should really pursue this as a business or a long-term thing, but I do know that I want to finish up X, Y and Z. It'll only take a month or two, and I think there's a good chance that it finally gets me over the hump.

This just kept happening over, and over, and over again. It's crazy.

And each time it happened, it felt like *this* was the time it *really would* only be another month or two. I was wrong last time, and the time before, and the time before, and the time before... but *this* time... this time I'll be right.

Wow. Articulating it like that really helps put things in perspective. I need to diagnose myself with a chronic case of the planning fallacy. I need to do a better job of adjusting in the opposite direction. I have a tendency to be *over*confident, so I need to adjust in the other direction, and be *less* confident. That's what you have to do with known biases: try to adjust in the other direction.

And with the planning fallacy in particular, there's a known cure: the outside view. Don't try to reason from the ground up. Look at how long similar things have taken in the past. Maybe use the reference class of "times I've thought it'd only be another month or two".

Man, that makes me laugh. "Times I've only thought it'd be another month or two." Ha! That reference class is full of miscalculations, so it's pretty clear that I need to adjust in the other direction pretty hard.

I say all of this stuff, but I still worry that I'm going to make the same mistakes again.

> It always takes longer than you expect, even when you take into account Hofstadter's Law.

Agility

Check out this excerpt from the first post in this series:

And I have this little voice in my head saying:

> Hey Adam... it's been over a year... you don't have *any* users. This like totally goes against the whole lean startup thing, y'know.

And then I say in response:

> I know, I know! But I really question whether the lean startup applies to this *particular* scenario. I *know* you want to avoid wasting time building things that people don't actually want. That's like the biggest startup sin there is. I know. But like, what am I supposed to do?
>
> My hypothesis is that once my app gets to the point where it's in the ballpark of Flopzilla or better, that people will want to use it. It takes time to get to that point. There isn't really a quick two week MVP I could build to test that hypothesis. I'm already trying to avoid building non-essential features, and focus on getting to that point as quickly as possible. So what am I supposed to do?
>
> If I released this and found that I had no users and no one wants this, what would that tell me? Just that people don't like *this* version of the app enough. Sure, ok. But the real question is whether they'll like the Ballpark Flopzilla version enough, or the Better Than Flopzillla version enough. My hypothesis is that they will, and releasing it now wouldn't invalidate that hypothesis. And I can't think of a way to test those hypotheses quickly. I think I just need to build it and see what happens.

I'm going to resist the temptation to respond to this right now. Right now I just want to tell the story. The story as it *actually* happened. But I do want to say that there were a lot of voices swimming around in my head questioning what I was doing.

I said in that post that I'm going to resist the temptation to respond to it. Here's where I *do* get to respond to it.

Here are my general thoughts about the whole lean startup thing:

1. I think the essence of it is to ruthlessly avoid spending time on things unnecessarily. Eg. you don't need to spend a week building a fancy navigation menu before you even know whether people want your product. I heartily, heartily support that message.

2. It can be tempting to think that there are no quick experiments you can run. This is usually wrong. Think more carefully, try to isolate the component assumptions, and get creative.

I definitely could have done a better job with (2).

3. But sometimes there *truly* aren't any quick experiments you can run. Eg. SpaceX. How is SpaceX supposed to build a quick MVP? (Well, there are some things they could do, but I don't want to get distracted from the central point that some hypotheses inherently just take a long time to validate.)

Now that I'm finished with Premium Poker Tools and am reflecting on what I've learned, it's time to add a fourth general thought.

4. If you *are* in a situation where (3) applies and are going to spend... I don't know... *two years and three months* testing a hypothesis... then the upside *damn well better be worth it!*

Pretty obvious. If the risk is big, the reward needs to also be big.

With Premium Poker Tools, the reward never big enough. It just wasn't. This is no SpaceX.

Even if my initial ideas about the market size were correct, and there was the potential to make $10-20M, that still isn't enough to justify spending so long testing a hypothesis.

However, I don't think it's that simple. This issue is very tangled up with the planning fallacy stuff I talked about in the "Just another month or two" section. I never actually *decided* to spend 2+ years on this.

Show me the money

https://www.youtube.com/watch?v=FFrag8ll85w (I had to)

I have another objection to the above section. I wasn't testing a hypothesis the whole time. No. About a year into it, I already *had* market validation. That's right. I already *had* validation.

I'm not sure when exactly I would say it occurred, but a big thing is when I posted to Reddit and someone asked where they can donate. Another big thing is seeing one of the most popular poker players in the world using my app. Another was having a bunch of people tell me that they would pay for it. Another is having random people email me thanking me for building the app and telling me how great it is.

Now, I know they say talk is cheap. I know how startup people always want to see *traction*. Users. Money. Growth. I didn't have any of that. But I did have all of the other stuff in the above paragraph! People said so many nice things to me. Sure, conventional wisdom might say that the above paragraph isn't enough, and that you need *real* traction...

But I'm a *Bayesian!* I'm better than that! Saying that you need *actual* traction is like how the scientific community waits too long before accepting something as true. Being a Bayesian, I can update in inches. I can be faster than science. I can be faster than conventional startup wisdom.

Well, that's how I used to think anyways. Now I run around my apartment with my fingers in my ears yelling SHOW ME THE MONEY!!!!!!!!!

I'm exaggerating of course. I don't actually do that. And I don't actually think you should ignore everything except "actual results". No, I'm still a Bayesian, and Bayesians don't throw away evidence. But given my experiences, I've come to believe that such evidence isn't *nearly* as strong as I had previously thought.

Deals fall through

This is similar to the above section. In the above section, I'm basically talking about how when people say "This app is awesome! I would pay for it!" it doesn't actually mean that much. This section is going to talk about how when potential business partners say "This is really interesting! Let's talk more!" or even "I'm in!", it doesn't actually mean that much.

You can read more about it in part one, but I've just had so many deals and stuff fall through. A lot of people were telling me that they were interested in working with me. Some even said that they *would* work with me. A verbal "yes". I thought to myself:

> Ok, great! This is pretty strong bayesian evidence right here. I wouldn't have so many people saying this stuff if they didn't mean it. Sure, maybe a few fall through, but not everyone.
>
> And also, I haven't even put myself out there too much. Just to throw out a number, maybe I've talked to 10% of the potential people who I could partner with. If I've gotten this much interest from the 10%, once I go after the 90%, I should end up with a good amount of partnerships.

Given my experiences, I've come to believe that verbal interest just isn't that telling. And this seems to mirror the experiences of others as well.

I don't want to say that verbal interest means nothing though. Honestly, I probably went too far in the above paragraph saying that it "just isn't that telling". I still have limited experience, and I'm not sure how strongly to weigh it as evidence.

But one thing does seem pretty clear: the lack of actual traction is stronger evidence than verbal interest. Eg. for me, I had a lot of people *saying* they're interested in partnering with me, but no one actually following through. I think the lack of follow through is stronger evidence than the verbal interest. Similarly, I had a lot of people *saying* they really like the app and stuff, but even when it was free I was only getting maybe 100 users/month, and they weren't spending that much time on the app. I should have paid more attention to that.

Build it and they'll come

I've always had a perspective that goes something like this:

> My app is at least in the same ballpark as Flopzilla. I feel pretty confident that it's a little bit better, actually. So then, I would think I should at least get 10-20% of the market, if not more.
>
> Yes, I know they were the first mover and have the brand recognition, but if my product is in the ballpark, I should still make a dent. I release it, people hear about it, some people like it and start using it, they tell their friends, people link to it in forums, put screenshots in blog posts, etc. I would think that if my product is as good as the market leader, through a process something like what I just described, I would get my slice of the pie.
>
> Furthermore, I would expect my slice to be proportionate to the quality of my product. If the product is a little worse than the competitors, maybe my slice is 5%. If it's a little better, maybe I get 30-40%. If it's way better, maybe it's 75%.
>
> And maybe that perspective is too optimistic. That's certainly possible. But it can't be *too* off the mark, right? Maybe if the product is a little worse I end up with 1% instead of 5%. Maybe if it's only a little better I get 5-10% instead of 30-40%. Maybe if it's way better I get 50% instead of 75%.

Although it's still a little counterintuitive to me, I have to say that my perspective is different now. The perspective I described above is some version of "build it and they'll come" (BIATC). I now thing that BIATC is pretty wrong.

I'm still not quite sure what the mechanism is, though. I suppose that it takes a lot to actually get word of mouth to happen. I suppose people don't actually do that much comparison shopping, and instead lean heavily towards things that are popular, have social proof, and that they stumble across organically, eg. in blogs.

I feel like I may have overestimated BIATC due the stuff I hear from YC folks. They talk a lot about focusing heavily on the product, as opposed to marketing it. I at least get the impression from them that making something people want is what it's all about, and that if you manage to do so, you'll have success. Here's Paul Graham in How to Start a Startup:

> It's not just startups that have to worry about this. I think most businesses that fail do it because they don't give customers what they want. Look at restaurants. A large percentage fail, about a quarter in the first year. But can you think of one restaurant that had really good food and went out of business?

> Restaurants with great food seem to prosper no matter what. A restaurant with great food can be expensive, crowded, noisy, dingy, out of the way, and even have bad service, and people will keep coming. It's true that a restaurant with mediocre food can sometimes attract customers through gimmicks. But that approach is very risky. It's more straightforward just to make the food good.

Using the example of restaurants, I actually can think of a lot of restaurants with great food that don't do so well. Hole-in-the-wall-type places with awesome food, but that are never too busy. And on the other hand, I can think of a lot of trendier restaurants that have terrible food but a lot of customers.

Still, I don't want to completely discount BIATC. I think it can be true in some situations. I just think those situations have to be pretty extreme. You need to 10x your competition. You need to be solving a hair on fire problem. You need to be building a painkiller, not a vitamin. You need your users to really, really love you. When you totally blow the competition away, or when you truly solve an important problem that hasn't been solved yet, yeah, when you build it, they'll come. Maybe that's what YC is trying to convey.

Customer acquisition is hard

Of course, I didn't actually just build my app and sit there expecting people to come to me. I did try to acquire customers.

It just didn't work. Despite the fact that I have a product that people say all of these nice things about, I only managed to get, let's say three paid users and 100 free trial sign ups. That amount of traction seems very much not in line with the quality of product I have, which makes me think that customer acquisition is very hard. Or maybe just that I'm still bad at it.

Here's what I tried, how it went, and what I learned:

Affiliate partnerships

This has always been Plan A. Find some people who already have huge followings eg. on YouTube, and piggyback off of that by offering them revenue share. Should be pretty simple. It's free money for them, and the product is good, so why wouldn't they want to do it? Especially when I ended up offering them 50%.

Well, I'm still not quite sure what the answer to that is.

Maybe they know how small the market is and how little money they'd make? Perhaps. But still, it's so low effort for them to throw a link in the description. And they're often already using poker software in their posts/books/videos, so it isn't any *extra* effort for them to use mine. In fact, mine is a better fit because mine is the only web app, and they could link to simulation results.

In talking to them, the response I usually get is that they've been meaning to check the app out but just haven't had time. This makes no sense to me because as poker professionals, I'd think that they are already spending time studying with software, so why not substitute mine in? Maybe they don't actually study? Maybe they don't want to spend the time messing around with a product they're unfamiliar with? Who knows.

My takeaway here is that if the deal you're offering people is only marginally beneficial, you might just end up with no partners.

YouTube and blogging

I think I have produced solid content with my YouTube channel and blog. People have told me that. But I've gotten only a minuscule amount of hits.

Again, my initial thinking was what I described in BIATC. That the amount of hits I get should be at least *roughly* proportional to the quality of my content.

Nope! That didn't happen! In retrospect, this makes sense. How are people supposed to discover your YouTube videos? YouTube recommends videos that are already popular, and chooses the videos that are already popular to put high in their search results. Similar with blogging. It's a chicken-egg problem that I have yet to figure out.

Paid advertising

There are a ton of huge companies built off of ad revenue: Google, Facebook, Instagram, Twitter, Reddit, etc. So then, that gives me the impression that paid ads are a huge thing that everyone is doing. And if they're doing it, they're doing it because it works for them, presumably. So paid ads have always been something that I assumed to work well.

Nope! The first place I learned this was in Julian Shapiro's guide. He said something that made it actually seem pretty clear in retrospect. If you *were* able to get an ad channel working profitably for you, you'd just be able to scale up with that channel, acquiring customers profitably, and making a ton of money. If every company were able to do that, entrepreneurship would be pretty easy.

Since reading his stuff, I started to hear other people say the same things, that it's really hard to do profitably and most companies never manage to do so. I still decided to give paid ads a shot, because it's a low risk high reward thing, but it turned out not to work for me.

Direct sales

I spent a pretty good amount of time doing direct sales. Some in person networking at a poker meetup, playing poker at the casinos, meeting people in coffee shops, Ubers, etc. Then I also spent some time DMing people on poker forums, and emailing poker players and coaches. I didn't go too crazy because I don't want to be spammy, but I definitely did some.

None of it really worked though. I think my takeaway mirrors the BIATC stuff, where you need to have something really valuable to actually get peoples attention.

I again found this surprising. I figured that when you actually are in a conversation with someone, they'd sort of give you the benefit of the doubt. But no, I didn't find that to be true.

Refer a friend

There was a point where I was offering $20 for every friend you refer. Pretty good, right?! Seemed pretty generous to me, and like something people would want to take advantage of. But no, that didn't happen. Still confusing to me, but I guess the lesson again is that it really takes a lot to get peoples attention.

Sharable simulation links

(I talk more about this in this post.)

Poker players spend a lot of time in forums discussing hands. They'll say things like, "you only have 33% equity, so you should fold". My app lets you link to a simulation that shows that you only have 33% equity. For non-users, it's in readonly mode. For users, when they could click the link and play around with the assumptions.

I thought this would be huge. Making it easier for people to discuss hands in the forums. But no, it did nothing. I think a big part of it is that most people in the forums approach it very casually and don't want to spend a lot of time on their comments. They just want to add their two cents, and then leave.

Free

The app was even *free* for a long period of time. I would have thought that this would attract a ton of users, but no, that didn't happen. More evidence that it takes a lot to actually get people to look in your direction, and that customer acquisition is hard.

People are lazy and irrational

Here's what I mean. It's really true that if you're a poker player and are actually trying to get better, that you should have some sort of poker software to study with. Coaches and professionals say this all the time. But people are still too lazy to do so.

Hell, they are just too lazy to study in general. I've said this before, but they prefer passive things. Watching a video and having someone "tell you the answers", even though that sort of passive study never works, regardless of the field. Eg. with math, you have to actually do a lot of practice problems yourself to get it to stick.

I've also heard a lot of coaches complain about students who pay them $100+/hr not do the homework that they are assigned. The coaches beg and plead, but the students just don't want to do the work.

That's the lazy part. I guess the irrational part plays into that, but what I really had in mind is that poker software is pretty easily a +ROI investment, but people still don't care. And if poker players aren't persuaded by +ROI investments, then I'm not sure what other demographic would be persuaded.

Everything else

There are definitely things that I'm forgetting. Hopefully I'll add to this post as I remember them.

If any readers out there have any advice, I'm all ears! I want to try to learn as much as I can from this.

Human instincts, symbol grounding, and the blank-slate neocortex

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

Intro: What is Common Cortical Algorithm (CCA) theory, and why does it matter for AGI?

As I discussed at Jeff Hawkins on neuromorphic AGI within 20 years, and was earlier discussed on LessWrong at The brain as a universal learning machine, there is a theory, due originally to Vernon Mountcastle in the 1970s, that the neocortex[1] (75% of the human brain by weight) consists of ~150,000 interconnected copies of a little module, the "cortical column", *each of which implements the same algorithm*. Following Jeff Hawkins, I'll call this the "common cortical algorithm" (CCA) theory. (I don't think that terminology is standard.)

So instead of saying that the human brain has a vision processing algorithm, motor control algorithm, language algorithm, planning algorithm, and so on, in CCA theory we say that (to a first approximation) we have a massive amount of "general-purpose neocortical tissue", and if you dump visual information into that tissue, it does visual processing, and if you connect that tissue to motor control pathways, it does motor control, etc.

Whether and to what extent CCA theory is true is, I think, very important for AGI forecasting, strategy, and both technical and non-technical safety research directions—see my answer here for more details.

Should we believe CCA theory?

CCA theory, as I'm using the term, is a simplified model. There are almost definitely a couple caveats to it:

1. There are sorta "hyperparameters" on the generic learning algorithm which seem to be set differently in different parts of the neocortex. For example, some areas of the cortex have higher or lower density of particular neuron types. There are other examples too.[2] I don't think this significantly undermines the usefulness or correctness of CCA theory, as long as these changes *really are* akin to hyperparameters, as opposed to specifying fundamentally different algorithms. So my reading of the evidence is that if you put, say, motor nerves coming out of visual cortex tissue, the tissue could do motor control, but it wouldn't do it quite as well as the motor cortex does.[3]
2. There is almost definitely a gross wiring diagram hardcoded in the genome—i.e., set of connections between different neocortical regions and each other, and other parts of the brain. These connections later get refined and edited during

> learning. Again, we can ask how much the existence of this innate gross wiring diagram undermines CCA theory. How complicated is the wiring diagram? Is it millions of connections among thousands of tiny regions, or just tens of connections among a few regions? Would the brain work at all if you started with a random wiring diagram? I don't know for sure, but for various reasons, my current belief is that this initial gross wiring diagram is *not* carrying much of the weight of human intelligence, and thus that this point is not a significant problem for the usefulness of CCA theory. (This is a loose statement; of course it depends on what questions you're asking.) I think of it more like: if it's biologically important to learn a concept space that's built out of associations between information sources X, Y, and Z, well, you just dump those three information streams into the same part of the cortex, and then the CCA will take it from there, and it will reliably build this concept space. So once you have the CCA nailed down, it kinda feels to me like you're *most* of the way there....[4]

Going beyond these caveats, I found pretty helpful literature reviews on both sides of the issue:

- The experimental evidence *for* CCA theory: see chapter 5 of *Rethinking Innateness* (1996)
- The experimental evidence *against* CCA theory: see chapter 5 of *The Blank Slate* by Steven Pinker (2002).

I won't go through the debate here, but after reading both of those I wound up feeling that CCA theory (with the caveats above) is probably right, though not 100% proven. Please comment if you've seen any other good references on this topic, especially more up-to-date ones.

(**Update:** I found another reference on CCA; see Gary Marcus vs Cortical Uniformity.)

CCA theory does not mean "no inductive biases"—of course there are inductive biases! It means that the inductive biases are sufficiently general and low-level that they work equally well for extremely diverse domains such as language, vision, motor control, planning, math homework, and so on. I typically think that the inductive biases are at a very low level, things like "we should model inputs using a certain type of data structure involving temporal sequences and spatial relations", and *not* higher-level semantic knowledge like intuitive biology or "when is it appropriate to feel guilty?" or tool use etc. (I don't even think object permanence or intuitive psychology are built into the neocortex; I think they're learned in early infancy. This is controversial and I won't try to justify it here. Well, intuitive psychology is a complicated case, see below.)

Anyway, that brings us to...

CCA theory vs human-universal traits and instincts

The main topic for this post is:

If Common Cortical Algorithm theory is true, then how do we account for all the human-universal instincts and behaviors that evolutionary psychologists talk about?

Indeed, we know that there are a diverse set of remarkably specific human instincts and mental behaviors evolved by natural selection. Again, Steven Pinker's *The Blank Slate* is a popularization of this argument; it ends with Donald E. Brown's giant list of "human universals", i.e. behaviors that are observed in every human culture.

Now, 75% of the human brain (by weight) is the neocortex, but the other 25% consists of various subcortical ("old-brain") structures like the amygdala, and these structures are perfectly capable of implementing specific instincts. But these structures do not have access to an intelligent world-model—only the neocortex does! So how can the brain implement instincts that require intelligent understanding? For example, maybe the fact that "Alice got two cookies and I only got one!" is represented in the neocortex as the activation of neural firing pattern 7482943. There's no obvious mechanism to connect this arbitrary, learned pattern to the "That's so unfair!!!" section of the amygdala. The neocortex doesn't know about unfairness, and the amygdala doesn't know about cookies. Quite a conundrum!

(**Update much later:** Throughout this post, wherever I wrote "amygdala", I should have said "hypothalamus and brainstem". See here for a better-informed discussion.)

This is really a symbol grounding problem, which is the other reason this post is relevant to AI alignment. When the human genome builds a human, it faces the same problem as a human programmer building an AI: how can one point a goal system at things in the world, when the internal representation of the world is a complicated, idiosyncratic, learned data structure? As we wrestle with the AI goal alignment problem, it's worth studying what human evolution did here.

List of ways that human-universal instincts and behaviors can exist despite CCA theory

Finally, the main part of this post. I don't know a complete answer, but here are some of the categories I've read about or thought of, and please comment on things I've left out or gotten wrong!

Mechanism 1: Simple hardcoded connections, not implemented in the neocortex

Example: Enjoying the taste of sweet things. This one is easy. I believe the nerve signals coming out of taste buds branch, with one branch going to the cortex to be integrated into the world model, and another branch going to subcortical regions. So the genes merely have to wire up the sweetness taste buds to the good-feelings subcortical regions.

Mechanism 2: Subcortex-supervised learning.

Example: Wanting to eat chocolate. This is different than the previous item because "sweet taste" refers to a specific innate physiological thing, whereas "chocolate" is a learned concept in the neocortex's world-model. So how do we learn to like chocolate? Because when we eat chocolate, we enjoy it (Mechanism 1 above). The neocortex learns to predict a sweet taste upon eating chocolate, and thus paints the world-model concept of chocolate with a "sweet taste" property. The supervisory

signal is multidimensional, such that the neocortex can learn to paint concepts with various labels like "painful", "disgusting", "comfortable", etc., and generate appropriate behaviors in response. (Vaguely related: the DeepMind paper Prefrontal cortex as a meta-reinforcement learning system.)

Mechanism 3: Same learning algorithm + same world = same internal model

Possible example: Intuitive biology. In *The Blank Slate* you can find a discussion of intuitive biology / essentialism, which "begins with the concept of an invisible essence residing in living things, which gives them their form and powers." Thus preschoolers will say that a dog altered to look like a cat is still a dog, yet a wooden toy boat cut into the shape of a toy car has in fact become a toy car. I think we can account for this very well by saying that everyone's neocortex has the same learning algorithm, and when they look at plants and animals they observe the same kinds of things, so we shouldn't be surprised that they wind up forming similar internal models and representations. I found a paper that tries to spell out how this works in more detail; I don't know if it's right, but it's interesting: free link, official link.

Mechansim 4: Human-universal memes

Example: Fire. I think this is pretty self-explanatory. People learn about fire from each other. No need to talk about neurons, beyond the more general issues of language and social learning discussed below.

Mechanism 5: "Two-process theory"

Possible example: Innate interest in human faces.[5] The subcortex-supervised learning mechanism above (Mechanism 2) can be thought of more broadly as an interaction between a hardwired subcortical system that creates a "ground truth", and a cortical learning algorithm that then learns to relate that ground truth to its complex internal representations. Here, Johnson's "two-process theory" for faces fits this same mold, but with a more complicated subcortical system for ground truth. In this theory, a subcortical system (ETA: specifically, the superior colliculus[6]) gets direct access to a low-resolution version of the visual field, and looks for a pattern with three blobs in locations corresponding to the eyes and mouth of a blurry face. When it finds such a pattern, it passes information to the cortex that this is a very important thing to attend to, and over time the cortex learns what faces actually look like (and suppresses the original subcortical template circuitry). Anyway, Johnson came up with this theory partly based on the observation that newborns are equally entranced by pictures of three blobs versus actual faces (each of which were much more interesting than other patterns), but after a few months the babies were more interested in actual face pictures than the three-blob pictures. (Not sure what Johnson would make of this twitter account.)

(Other possible examples of instincts formed by two-process theory: fear of snakes, interest in human speech sounds, sexual attraction.)

(Update: See my later post Inner alignment in the brain for a more fleshed-out discussion of this mechanism.)

Mechanism 6: Time-windows

Examples: Filial imprinting in animals, incest repulsion (Westermarck effect) in humans. Filial imprinting is a famous result where newborn chicks (and many other species) form a permanent attachment to the most conspicuous moving object

that they see in a certain period shortly after hatching. In nature, they always imprint on their mother, but in lab experiments, chicks can be made to imprint on a person, or even a box. As with other mechanisms here, time-windows provides a nice solution to the symbol grounding problem, in that the genes don't need to know what precise collection of neurons corresponds to "mother", they only need to set up a time window and a way to point to "conspicuous moving objects", which is presumably easier. The brain mechanism of filial imprinting has been studied in detail for chicks, and consists of the combination of time-windows plus the two-process model (mechanism 5 above). In fact, I think the two-process model was proven in chick brains before it was postulated in human brains.

There likewise seem to be various time-window effects in people, such as the Westermarck effect, a sexual repulsion between two people raised together as young children (an instinct which presumably evolved to reduce incest).

Mechanism 7 (speculative): empathetic grounding of intuitive psychology.

Possible example: Social emotions (gratitude, sympathy, guilt,...) Again, the problem is that the neocortex is the only place with enough information to, say, decide when someone slighted you, so there's no "ground truth" to use for subcortex-supervised learning. At first I was thinking that the two-process model for human faces and speech could be playing a role, but as far as I know, deaf-blind people have the normal suite of social emotions, so that's not it either. I looked in the literature a bit and couldn't find anything helpful. So, I made up this possible mechanism (*warning: wild speculation*).

Step 1 is that a baby's neocortex builds a "predicting my own emotions" model using normal subcortex-supervised learning (Mechanism 2 above). Then a normal Hebbian learning mechanism makes two-way connections between the relevant subcortical structures (amygdala) and the cortical neurons involved in this predictive model.

Step 2 is that the neocortex's universal learning algorithm will, in the normal course of development, naturally discover that this same "predicting my own emotions" model from step 1 can be reused to predict other people's emotions (cf. Mechanism 3 above), forming the basis for intuitive psychology. Now, because of those connections-to-the-amygdala mentioned in step 1, the amygdala is incidentally getting signals from the neocortex when the latter predicts that *someone else* is angry, for example.

Step 3 is that the amygdala (and/or neocortex) somehow learns the difference between the intuitive psychology model running in first-person mode versus empathetic mode, and can thus generate appropriate reactions, with one pathway for "being angry" and a different pathway for "knowing that someone else is angry".

So let's now return to my cookie puzzle above. Alice gets two cookies and I only get one. How can I feel it's unfair, given that the neocortex doesn't have a built-in notion of unfairness, and the amygdala doesn't know what cookies are? The answer would be: thanks to subcortex-supervised learning, the amygdala gets a message that one yummy cookie is coming, but the neocortex also thinks "Alice is even happier", and that thought also recruits the amygdala, since intuitive psychology is built on empathetic modeling. Now the amygdala knows that I'm gonna get something good, but that Alice is gonna get something even better, and that combination (in the current emotional context) triggers the amygdala to send out waves of jealousy and indignation. This is then a new supervisory signal for the neocortex, which allows the neocortex to gradually develop a model of fairness, which in turn feeds back into the

intuitive psychology module, and thereby back to the amygdala, allowing the amygdala to execute more complicated innate emotional responses in the future, and so on.

(Update: See my later post Inner alignment in the brain for a slightly more fleshed-out discussion of this mechanism.)

The special case of language.

It's tempting to put language in the category of memes (mechanism 4 above)—we do generally learn language from each other—but it's not really, because apparently groups of kids can invent grammatical languages from scratch (e.g. Nicaraguan Sign Language). My current guess is that it combines three things: (1) a two-process mechanism (Mechanism 5 above) that makes people highly attentive to human speech sounds. (2) possibly "hyperparameter tuning" in the language-learning areas of the cortex, e.g. maybe to support taller compositional hierarchies than would be required elsewhere in the cortex. (3) The fact that language can sculpt itself to the common cortical algorithm rather than the other way around—i.e., *maybe* "*grammatical language*" *is just another word for* "*a language that conforms to the types of representations and data structures that are natively supported by the common cortical algorithm*".

By the way, lots of people (including Steven Pinker) seem to argue that language processing is a fundamentally different and harder task than, say, visual processing, because language requires symbolic representations, composition, recursion, etc. I don't understand this argument; I think vision processing needs the exact same things! I don't see a fundamental difference between the visual-processing system knowing that "this sheet of paper is part of my notebook", and the grammatical "this prepositional phrase is part of this noun phrase". Likewise, I don't see a difference between recognizing a background object interrupted by a foreground occlusion, versus recognizing a noun phrase interrupted by an interjection. It seems to me like a similar set of problems and solutions, which again strengthens my belief in CCA theory.

Conclusion

When I initially read about CCA theory, I didn't take it too seriously because I didn't see how instincts could be compatible with it. But I now find it pretty likely that there's no fundamental incompatibility. So having removed that obstacle, and also read the literature a bit more, I'm much more inclined to believe that CCA theory is fundamentally correct.

Again, I'm learning as I go, and in some cases making things up as I go along. Please share any thoughts and pointers!

---

1. I'll be talking a lot about the neocortex in this article, but shout-out to the thalamus and hippocampus, the other two primary parts of the brain's predictive-world-model-building-system. I'm just leaving them out for simplicity; this doesn't have any important implications for this article. ↩

2. More examples of region-to-region variation in the neocortex that are (plausibly) genetically-coded: (1) Spindle neurons only exist in a couple specific parts of the

neocortex. I don't really know what's the deal with those. Kurzweil claims they're important for social emotions and empathy, if I recall correctly. Hmmm. (2) "Sensitive windows" (see Dehaene): Low-level sensory processing areas more-or-less lock themselves down to prevent further learning very early in life, and certain language-processing areas lock themselves down somewhat later, and high-level conceptual areas don't ever lock themselves down at all (at least, not as completely). I bet that's genetically hardwired. I guess psychedelics can undermine this lock-down mechanism? ↩

3. I have heard that the primary motor cortex is not the only part of the neocortex that emits motor commands, but don't know the details. ↩

4. Also, people who lose various parts of the neocortex are often capable of full recovery, if it happens early enough in infancy, which suggests to me that the CCA's wiring-via-learning capability is doing most of the work, and maybe the innate wiring diagram is mostly just getting things set up more quickly and reliably. ↩

5. See *Rethinking Innateness* p116, or better yet Johnson's article ↩

6. See, for example, Fast Detector/First Responder: Interactions between the Superior Colliculus-Pulvinar Pathway and Stimuli Relevant to Primates. Also, let us pause and reflect on the fact that humans have two different visual processing systems! Pretty cool! The most famous consequence is blindsight, a condition where the ~~subconscious~~ midbrain vision processing system (superior colliculus) is intact but the ~~conscious~~ neocortical visual processing system is not working. This study proves that blindsighted people can recognize not just faces but specific facial expressions. I strongly suspect blindsighted people would react to snakes and spiders too, but can't find any good studies (that study in the previous sentence regrettably used stationary pictures of spiders and snakes, not videos of them scampering and slithering). ↩

What are your strategies for avoiding micro-mistakes?

I've recently been spending more time doing things that involve algebra and/or symbol manipulation (after a while not doing these things by hand that often) and have noticed that small mistakes cost me a lot of time. Specifically, I can usually catch such mistakes by double-checking my work, but the cost of not being able to trust my initial results and redo steps is very high. High enough that I'm willing to spend time working to reduce the number of such mistakes I make even if it means slowing down quite a bit or adopting some other costly process.

If you've either developed such strategies for avoiding making such mistakes or would good at it in the first place, what do you do?

Two notes on the type of answers I'm looking for:

1. I should note that one answer is just to use something like WolframAlpha or Mathematica, which I do. That said, I'm still interested in not having to rely on such tools for things in the general symbol manipulation reference class as I don't like relying on my computer being present to do these sorts of things.

2. I did do some looking around for work addressing this (found this for example), but most of it suggested basic strategies that I already implement like being neat and checking your work.