Best of LessWrong: October 2018

1. Embedded Agents
2. Some cruxes on impactful alternatives to AI policy work
3. The Rocket Alignment Problem
4. Public Positions and Private Guts
5. Good Samaritans in experiments
6. The funnel of human experience
7. Coordination Problems in Evolution: Eigen's Paradox
8. Being a Robust Agent
9. In praise of heuristics
10. Debate Rules In Benjamin Franklin's Junto
11. On Doing the Improbable
12. List of previous prediction market projects
13. Genomic Prediction is now offering embryo selection
14. Book review: The Complacent Class
15. An Undergraduate Reading Of: Semantic information, autonomous agency and non-equilibrium statistical physics
16. Fasting Mimicking Diet Looks Pretty Good
17. Alignment Newsletter #28
18. What will the long-term future of employment look like?
19. The Kelly Criterion
20. A Rationality Condition for CDT Is That It Equal EDT (Part 2)
21. Two Kinds of Technology Change
22. Population Aging as an Impediment to Addressing Global Catastrophic Risks
23. The Art of the Overbet
24. Introducing the AI Alignment Forum (FAQ)
25. You're never wrong injecting complexity, but rarely you're right
26. Computerphile discusses MIRI's "Logical Induction" paper
27. Preface to the sequence on value learning
28. Reflections on Being 30
29. Stop buttons and causal graphs
30. Mark Eichenlaub: How to develop scientific intuition
31. Alignment Newsletter #27
32. Alignment Newsletter #29
33. Why do we like stories?
34. Maps of Meaning: Abridged and Translated
35. The Valley of Bad Theory
36. Things I Learned From Working With A Marketing Advisor
37. Where is my Flying Car?
38. A compendium of conundrums
39. Feedback from emotions
40. Effective Altruism Book Review: Radical Abundance (Nanotechnology)
41. Why don’t we treat geniuses like professional athletes?
42. Deconstructing Biases In Media
43. New /r/gwern subreddit for link-sharing
44. On insecurity as a friend
45. Kalman Filter for Bayesians
46. Thoughts on short timelines
47. Coordination Problems in Evolution: The Rise of Eukaryotes
48. Cognitive Enhancers: Mechanisms And Tradeoffs
49. Epistemic Spot Check: The Dorito Effect (Mark Schatzker)
50. Decision Theory Anti-realism

Best of LessWrong: October 2018

1. Embedded Agents
2. Some cruxes on impactful alternatives to AI policy work
3. The Rocket Alignment Problem
4. Public Positions and Private Guts
5. Good Samaritans in experiments
6. The funnel of human experience
7. Coordination Problems in Evolution: Eigen's Paradox
8. Being a Robust Agent
9. In praise of heuristics
10. Debate Rules In Benjamin Franklin's Junto
11. On Doing the Improbable
12. List of previous prediction market projects
13. Genomic Prediction is now offering embryo selection
14. Book review: The Complacent Class
15. An Undergraduate Reading Of: Semantic information, autonomous agency and non-equilibrium statistical physics
16. Fasting Mimicking Diet Looks Pretty Good
17. Alignment Newsletter #28
18. What will the long-term future of employment look like?
19. The Kelly Criterion
20. A Rationality Condition for CDT Is That It Equal EDT (Part 2)
21. Two Kinds of Technology Change
22. Population Aging as an Impediment to Addressing Global Catastrophic Risks
23. The Art of the Overbet
24. Introducing the AI Alignment Forum (FAQ)
25. You're never wrong injecting complexity, but rarely you're right
26. Computerphile discusses MIRI's "Logical Induction" paper
27. Preface to the sequence on value learning
28. Reflections on Being 30
29. Stop buttons and causal graphs
30. Mark Eichenlaub: How to develop scientific intuition
31. Alignment Newsletter #27
32. Alignment Newsletter #29
33. Why do we like stories?
34. Maps of Meaning: Abridged and Translated
35. The Valley of Bad Theory
36. Things I Learned From Working With A Marketing Advisor
37. Where is my Flying Car?
38. A compendium of conundrums
39. Feedback from emotions
40. Effective Altruism Book Review: Radical Abundance (Nanotechnology)
41. Why don’t we treat geniuses like professional athletes?
42. Deconstructing Biases In Media
43. New /r/gwern subreddit for link-sharing
44. On insecurity as a friend
45. Kalman Filter for Bayesians
46. Thoughts on short timelines
47. Coordination Problems in Evolution: The Rise of Eukaryotes
48. Cognitive Enhancers: Mechanisms And Tradeoffs

49. Epistemic Spot Check: The Dorito Effect (Mark Schatzker)
50. Decision Theory Anti-realism

Embedded Agents

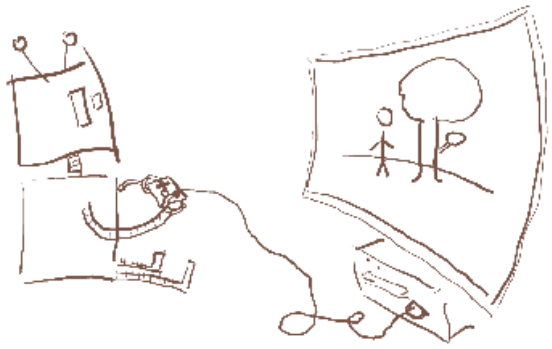
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(A longer text-based version of this post is also available on MIRI's blog here, and the bibliography for the whole sequence can be found here)

Embedded Agency

Abram Demski & Scott Garrabrant

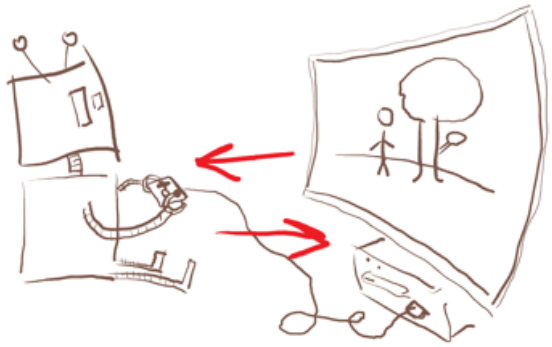
This is Alexei.



Alexei is playing a

video game.

Alexei interacts with the environment via well-defined i/o channels.

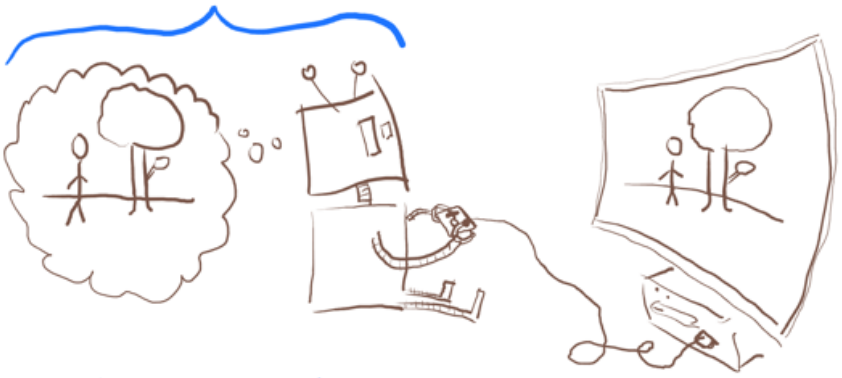


This means Alexei has a clearly-defined functional relationship with the environment, defining action consequences.

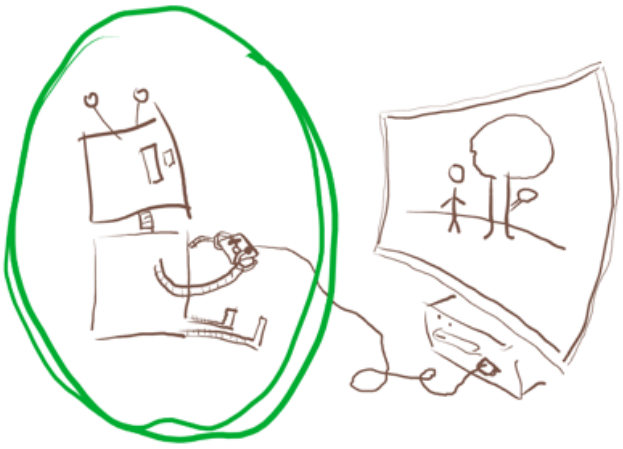
Alexei can hold the entire environment in mind.

Alexei may need to learn what the environment is like, but in doing so, can represent every detail.

Alexei thinks about manipulating and controlling the environment, but not himself.

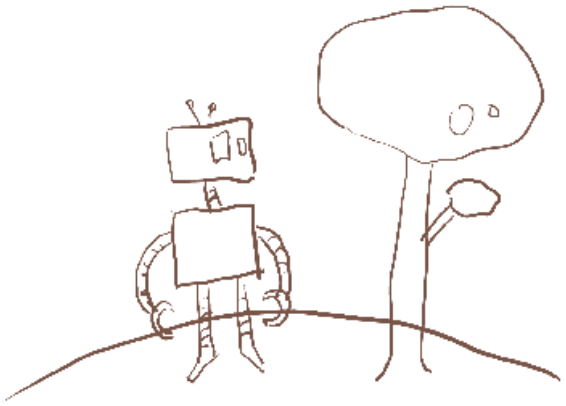


He doesn't have the opportunity or risk of arbitrary self-modification, because the environment can't really touch him. He can't really die, either; he only has to worry about restarts.



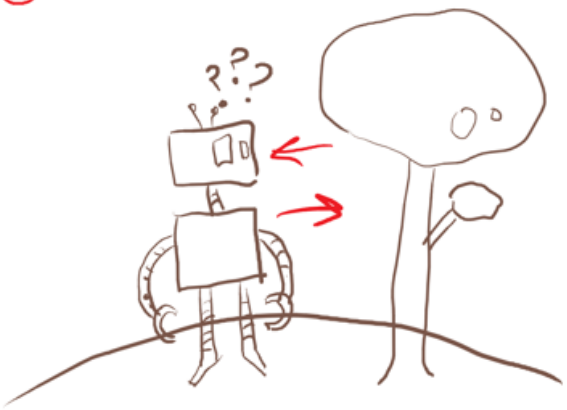
Alexei may think like a reductive scientist about the world, breaking it into parts, but the concept of agent is non-reductive; Alexei is an indivisible atom which produces actions.

This is Emmy.



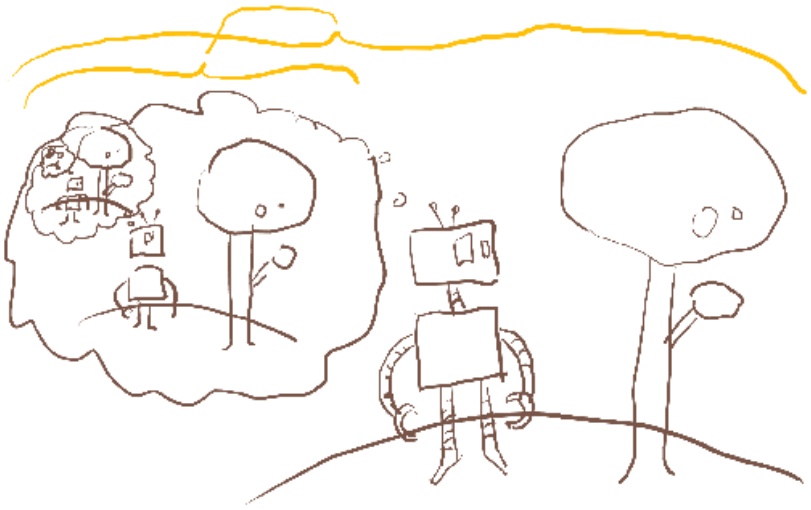
Emmy is playing real life.

Emmy is part of the universe, not sitting outside, so it is hard for her to imagine taking "different actions".



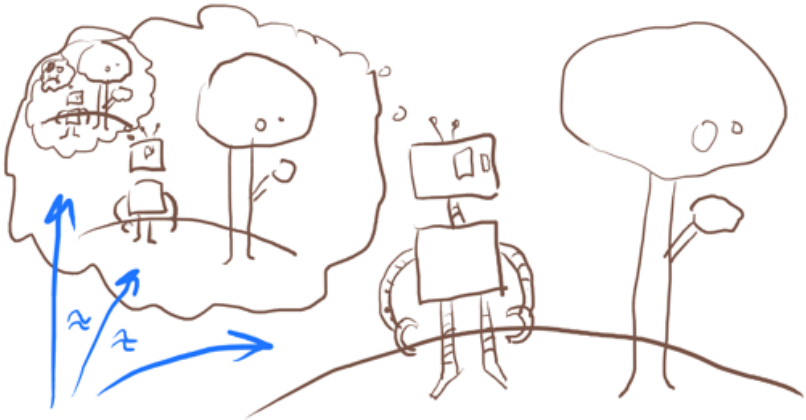
Alexei can poke the universe and see what happens. Emmy is the universe poking itself.

Emmy can't hold the entire world in her head.



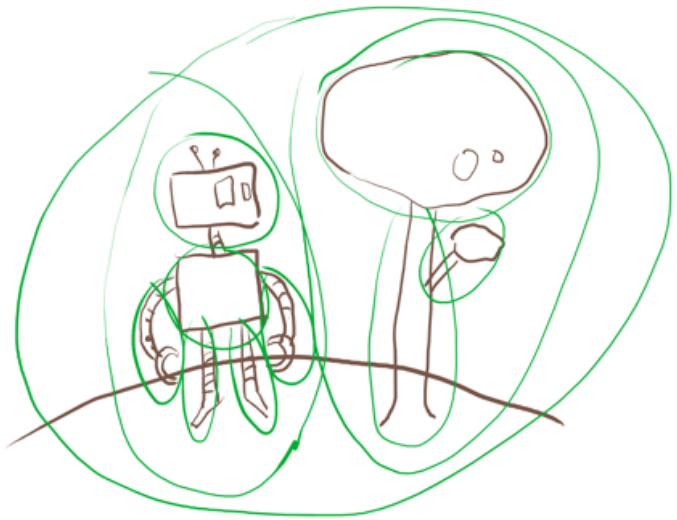
Any model she uses will be very partial and approximate.

Emmy can reflect and self-improve.



Emmy thinks about how to think about how to win, where Alexei only thinks about how to win.

Emmy is made of parts, just like everything else.

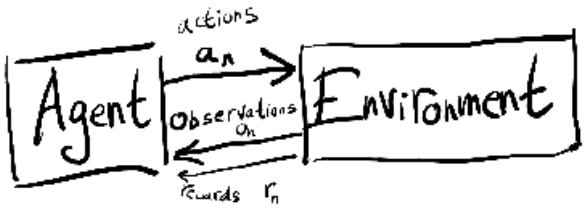


She isn't really a unitary entity; she is just a bunch of stuff. Somehow we get an agent out of non-agentic pieces.

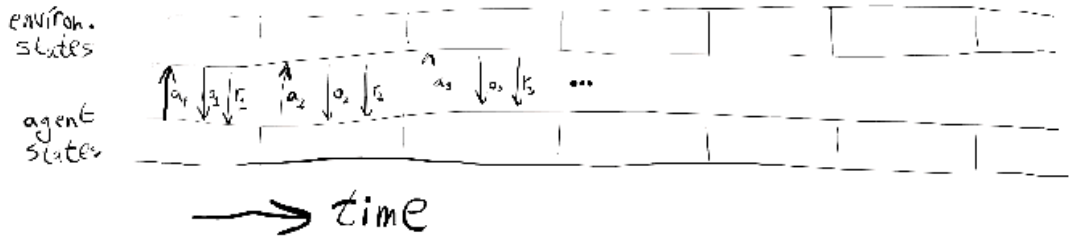
This is Marcus Hutter.

$$a_k = \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q:U(q,a_1 \ldots a_m) = o_1 r_1 \ldots o_m r_m} 2^{-l(q)}$$

Marcus Hutter's AIXI model tells us all about the kind of thinking Alexei needs to do to be good at winning video games.



The AIXI model sets up a somewhat symmetric relationship between the agent and environment: the agent produces a sequence of actions which are a function of previous observations and rewards, while the environment produces observations and rewards in a way which is a function of the previous actions.



AIXI doesn't know which environment it is interacting with, so it uses a probability distribution on all computable environments.

Its task is to learn about its environment through observation, and plan to get as much reward as possible (taking into account its uncertainty to hedge its bets).

Agent models like AIXI are dualistic: the agent exists outside of the environment, with only set interactions between agent-stuff and environment-stuff. They require the agent to be larger than the environment and don't tend to model self-referential reasoning because the agent is made of different stuff than what the agent reasons about.

These dualistic assumptions are not unique to AIXI; they are very common among models of rational agency.

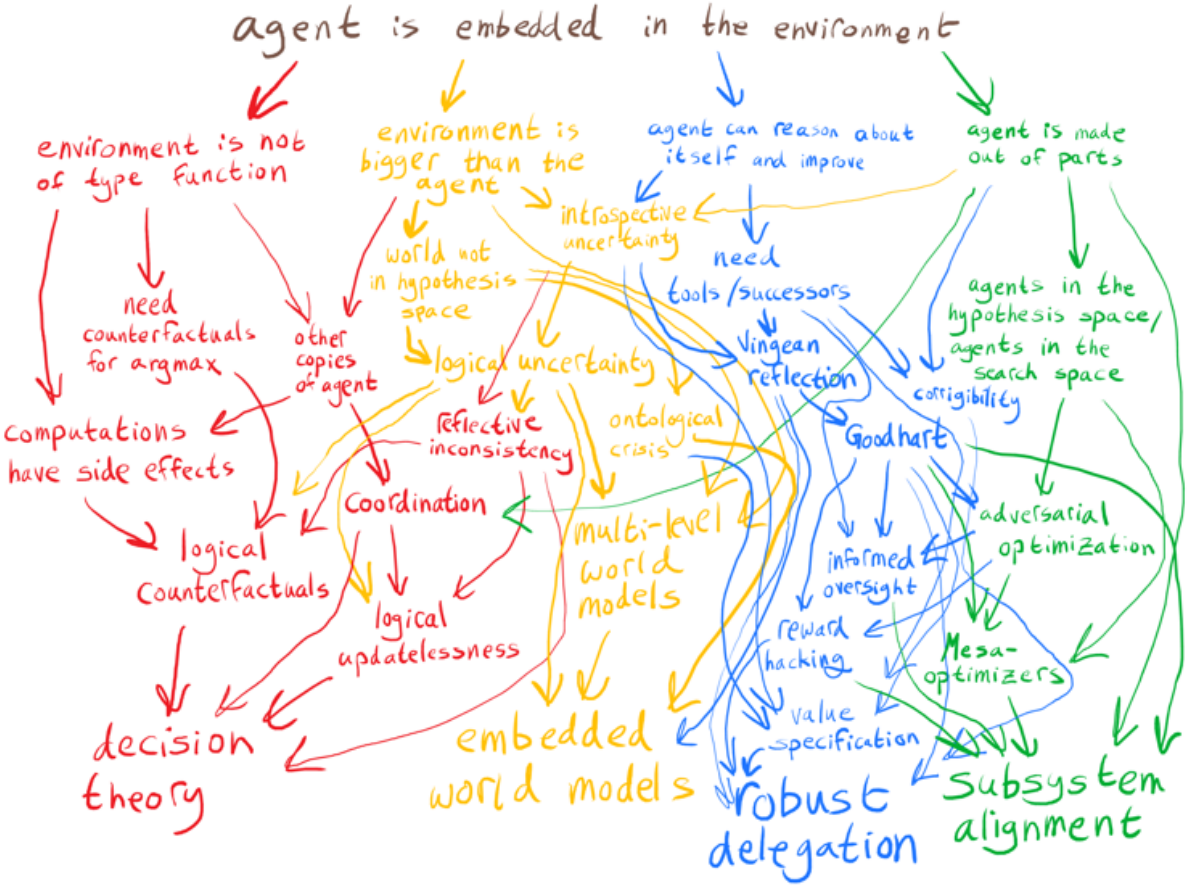
We would like to understand agents like Emmy as well as we currently understand agents like Alexei. AIXI serves as an illustration of what this high level of understanding looks like.

However, agents in the real world are forced to break important assumptions of the AIXI model.

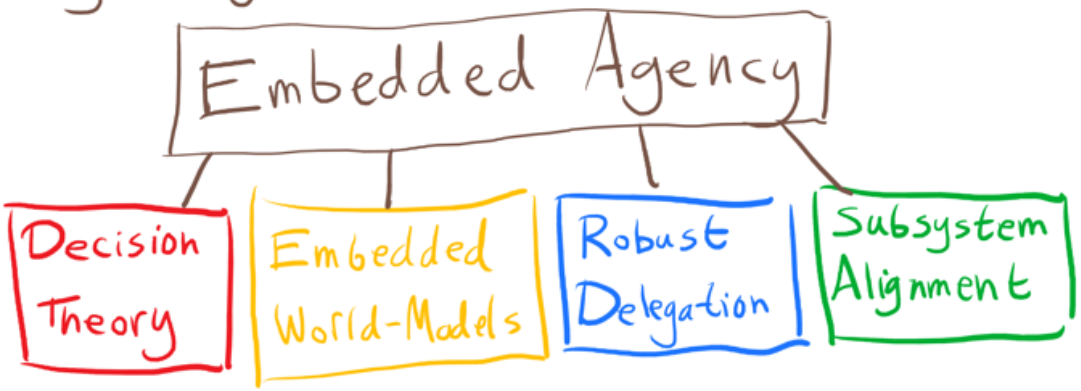
Agents like Emmy are embedded in their environments. Embedded agents break the dualistic assumptions:

- → not given well-defined i/o channels
- → smaller than the environment
- → able to reason about themselves & self-improve
- → made of parts like the environment

The four properties don't cleanly divide everything up. They're more like four ways of looking at the same problem than mutually exclusive categories.


agent is embedded in the environment
environment is not of type function
environment is bigger than the agent
agent can reason about itself and improve
agent is made out of parts
need counterfactuals for argmax
other copies of agent
world not in hypothesis space
introspective uncertainty
need tools/successors
logical uncertainty
Vingean reflection
agents in the hypothesis space/ agents in the search space
corrigibility
computations have side effects
reflective inconsistency
ontological crisis
Goodhart
Coordination
multi-level world models
adversarial optimization
informed oversight
logical counterfactuals
logical updatelessness
reward hacking
Mesa-optimizers
value specification
decision theory
embedded world models
robust delegation
subsystem alignment

We can cluster problems of embedded agency into four subfields, which is as close to a nice partitioning as we're likely to get.



These have a rough correspondence to the four properties of embedded agents.

Decision Theory: Adapting classical decision theory to embedded agents

- → Counterfactuals
- → Newcomblike problems; copies of the agent
- → Reasoning about other agents
- → Extortion problems
- → Coordination problems
- → Logical counterfactuals
- → Logical updatelessness

Embedded World-Models: understanding epistemic states appropriate for embedded agents

- → world not in hypothesis space ("realizability"/"grain of truth" problem)
- → logical uncertainty
- → high-level models
- → multi-level models
- → ontological crises
- → agent must be in world-model (Naturalized Induction)
- → anthropic reasoning

Robust Delegation: understanding what trust relationships can exist between an agent and its future self or other agents it can delegate to.

- → Vingean reflection
- → Tiling problem
- → Averting Goodhart's Law
- → Value Loading
- → Corrigibility
- → Informed Oversight

Subsystem Alignment: Ensuring that subsystems are not working at cross purposes; avoiding subprocesses optimizing for unintended goals.

- → Benign Induction
- → Benign Optimization
- → Transparency
- → Mesa-Optimizers

These four areas point to our biggest confusions about how highly intelligent systems could work in the real world.

The next four sections will go into these areas in more detail.

Some cruxes on impactful alternatives to AI policy work

Ben Pace and I (Richard Ngo) recently did a public double crux at the Berkeley REACH on how valuable it is for people to go into AI policy and strategy work: I was optimistic and Ben was pessimistic. During the actual event, we didn't come anywhere near to finding a double crux on that issue. But after a lot of subsequent discussion, we've come up with some more general cruxes about where impact comes from.

I found Ben's model of how to have impact very interesting, and so in this post I've tried to explain it, along with my disagreements. Ben liked the goal of writing up a rough summary of our positions and having further discussion in the comments, so while he edited it somewhat he doesn't at all think that it's a perfect argument, and it's not what he'd write if he spent 10 hours on it. He endorsed the wording of the cruxes as broadly accurate.

(During the double crux, we also discussed how the heavy-tailed worldview applies to community building, but decided on this post to focus on the object level of what impact looks like.)

Note from Ben: "I am not an expert in policy, and have not put more than about 20-30 hours of thought into it total as a career path. But, as I recently heard Robin Hanson say, there's a common situation that looks like this: some people have a shiny idea that they think about a great deal and work through the details of, that folks in other areas are skeptical of given their particular models of how the world works. Even though the skeptics have less detail, it can be useful to publicly say precisely why they're skeptical.

In this case I'm often skeptical when folks tell me they're working to reduce x-risk by focusing on policy. Folks doing policy work in AI might be right, and I might be wrong, but it seemed like a good use of time to start a discussion with Richard about how I was thinking about it and what would change my mind. If the following discussion causes me to change my mind on this question, I'll be really super happy with it."

Ben's model: Life in a heavy-tailed world

A heavy-tailed distribution is one where the probability of extreme outcomes doesn't drop very rapidly, meaning that outliers therefore dominate the expectation of the distribution. Owen Cotton-Barratt has written a brief explanation of the idea here. Examples of heavy-tailed distributions include the Pareto distribution and the log-normal distribution; other phrases people use to point at this concept include 'power laws' (see Zero to One) and 'black swans' (see the recent SSC book review). Wealth is a heavy-tailed distribution, because many people are clustered relatively near the median, but the wealthiest people are millions of times further away. Human height and weight and running speed are not heavy-tailed; there is no man as tall as 100 people.

There are three key claims that make up Ben's view.

The first claim is that, since the industrial revolution, we live in a world where the impact that small groups can have is much more heavy-tailed than in the past.

- People can affect incredibly large numbers of other people worldwide. The Internet is an example of a revolutionary development which allows this to happen very quickly.
- Startups are becoming unicorns unprecedentedly quickly, and their valuations are very heavily skewed.
- The impact of global health interventions is heavy-tail distributed. So is funding raised by Effective Altruism - two donors have contributed more money than everyone else combined.
- Google and Wikipedia qualitatively changed how people access knowledge; people don't need to argue about verifiable facts any more.
- Facebook qualitatively changed how people interact with each other (e.g. FB events is a crucial tool for most local EA groups), and can swing elections.
- It's not just that we got more extreme versions of the same things, but rather that we can get unforeseen types of outcomes.
- The books *HPMOR* and *Superintelligence* both led to mass changes in plans towards more effective ends via the efforts of individuals and small groups.

The second claim is that you should put significant effort into re-orienting yourself to use high-variance strategies.

- Ben thinks that recommending strategies which are *safe* and *low-risk* is insane when pulling out of a heavy-tailed distribution. You want everyone to be taking high-variance strategies.
    - This is only true if the tails are long to the right and not to the left, which seems true to Ben. Most projects tend to end up not pulling any useful levers whatever and just do nothing, but a few pull crucial levers and solve open problems or increase capacity for coordination.
- Your intuitions were built for the ancestral environment where you didn't need to be able to think about coordinating humans on the scale of millions or billions, and yet you still rely heavily on the intuitions you're built with in navigating the modern environment.
- Scope insensitivity, framing effects, taboo tradeoffs, and risk aversion, are the key things here. You need to learn to train your S1 to understand *math*.
    - By default, you're not going to spend enough effort finding or executing high-variance strategies.
- We're still only 20 years into the internet era. Things keep changing qualitatively, but Ben feels like everyone keeps adjusting to the new technology as if it were always this way.
- Ben: "My straw model of the vast majority of people's attitudes is: I guess Facebook and Twitter are just things now. I won't spend time thinking about whether I could build a platform as successful as those two but optimised better for e.g. intellectual progress or social coordination - basically not just money."
- Ben: "I do note that never in history has change been happening so quickly, so it makes sense that people's intuitions are off."
- While many institutions have been redesigned to fit the internet, Ben feels like almost nobody is trying to improve institutions like science on a large scale, and that this is clear low-hanging altruistic fruit.
- The Open Philanthropy Project has gone through this process of updating away from safe, low-risk bets with GiveWell, toward hits-based giving, which is an example of this kind of move.

The third claim is that AI policy is not a good place to get big wins nor to learn the relevant mindset.

- Ben: “On a first glance, governments, politics and policy looks like the sort of place where I would not expect to find highly exploitable strategies, nor a place that will teach me the sorts of thinking that will help me find them in future.”
- People in policy spend a lot of time thinking about how to influence governments. But governments are generally too conventional and slow to reap the benefits of weird actions with extreme outcomes.
- Working in policy doesn't cultivate the right type of thinking. You're usually in a conventional governmental (or academic) environment, stuck inside the system, getting seduced by local incentive gradients and prestige hierarchies. You often need to spend a long time working your way to positions of actual importance in the government, which leaves you prone to value drift or over-specialisation in the wrong thing.
    - At the very least, you have to operate on the local incentives as well as someone who actually cares about them, which can be damaging to one’s ability to think clearly.
- Political landscapes are not the sort of environment where people can easily ignore the local social incentives to focus on long-term, global goals. Short term thinking (election cycles, media coverage, etc) is not the sort of thinking that lets you build a new institution over 10 years or more.
    - Ben: “When I’ve talked to senior political people, I’ve often heard things of the sort ‘We were working on a big strategy to improve infrastructure / international aid / tech policy etc, but then suddenly public approval changed and then we couldn’t make headway / our party wasn’t in power / etc.’ which makes me think long term planning is strongly disincentivised.”
- One lesson of a heavy-tailed world is that signals that you’re taking safe bets are *anti-signals* of value. Many people following a standard academic track saying “Yeah, I’m gonna get a masters in public policy” sounds *fine*, *sensible, and safe*, and therefore *cannot* be an active sign that you will do something a million times more impactful than the median.

The above is not a full, gears-level analysis of how to find and exploit a heavy tail, because almost all of the work here lies in identifying the particular strategy. Nevertheless, because of the considerations above, Ben thinks that talented, agenty and rational people should be able in many cases to identify places to win, and then execute those plans, and that this is much less the case in policy.

Richard's model: Business (mostly) as usual

I disagree with Ben on all three points above, to varying degrees.

On the first point, I agree that the distribution of success has become much more heavy-tailed since the industrial revolution. However, I think the distribution of success is often very different from the distribution of impact, because of replacement effects. If Facebook hadn't become the leading social network, then MySpace would have. If not Google, then Yahoo. If not Newton, then Leibniz (and if Newton, then Leibniz anyway). Probably the alternatives would have been somewhat worse, but not significantly so (and if they were, different competitors would have come along). The distinguishing trait of modernity is that even a small difference in quality can lead to a huge difference in earnings, via network effects and global markets. But that isn't

particularly interesting from an x-risk perspective, because money isn't anywhere near being our main bottleneck.

You might think that since Facebook has billions of users, their executives are a small group with a huge amount of power, but I claim that they're much more constrained by competitive pressures than they seem. Their success depends on the loyalty of their users, but the bigger they are, the easier it is for them to seem untrustworthy. They also need to be particularly careful since antitrust cases have busted the dominance of several massive tech companies before. (While they could swing a few elections before being heavily punished, I don't think this is unique to the internet age - a small cabal of newspaper owners could probably have done the same centuries ago). Similarly, I think the founders of Wikipedia actually had fairly little counterfactual impact, and currently have fairly little power, because they're reliant on editors who are committed to impartiality.

What we should be more interested in is cases where small groups didn't just ride a trend, but actually created or significantly boosted it. Even in those cases, though, there's a big difference between success and impact. Lots of people have become very rich from shuffling around financial products or ad space in novel ways. But if we look at the last fifty years overall, they're far from dominated by extreme transformative events - in fact, Western societies have changed very little in most ways. Apart from IT, our technology remains roughly the same, our physical surroundings are pretty similar, and our standards of living have stayed flat or even dropped slightly. (This is a version of Tyler Cowen and Peter Thiel's views; for a better articulation, I recommend *The Great Stagnation* or *The Complacent Class).* Well, isn't IT enough to make up for that? I think it will be eventually, as AI develops, but right now most of the time spent on the internet is wasted. I don't think current IT has had much of an effect by standard metrics of labour productivity, for example.

Should you pivot?

Ben might claim that this is because few people have been optimising hard for positive impact using high-variance strategies. While I agree to some extent, I also think that there are pretty strong incentives to have impact regardless. We're in the sort of startup economy where scale comes first and monetisation comes second, and so entrepreneurs already strive to create products which influence millions of people even when there's no clear way to profit from them. And entrepreneurs are definitely no strangers to high-variance strategies, so I expect most approaches to large-scale influence to already have been tried.

On the other hand, I do think that reducing existential risk is an area where a small group of people are managing to have a large influence, a claim which seems to contrast with the assertion above. I'm not entirely sure how to resolve this tension, but I've been thinking lately about an analogy from finance. Here's Tyler Cowen:

> I see a lot of money managers, so there's Ray Dalio at Bridgewater. He saw one basic point about real interest rates, made billions off of that over a great run. Now it's not obvious he and his team knew any better than anyone else.
>
> Peter Lynch, he had fantastic insights into consumer products. Use stuff, see how you like it, buy that stock. He believed that in an age when consumer product stocks were taking off.
>
> Warren Buffett, a certain kind of value investing. Worked great for a while, no big success, a lot of big failures in recent times.

The analogy isn’t perfect, but the idea I want to extract is something like: once you’ve identified a winning strategy or idea, you can achieve great things by exploiting it - but this shouldn’t be taken as strong evidence that you can do exceptional things in general. For example, having a certain type of personality and being a fan of science fiction is very useful in identifying x-risk as a priority, but not very useful in founding a successful startup. Similarly, being a philosopher is very useful in identifying that helping the global poor is morally important, but not very useful in figuring out how to solve systemic poverty.

From this mindset, instead of looking for big wins like “improving intellectual coordination”, we should be looking for things which are easy conditional on existential risk actually being important, and conditional on the particular skillsets of x-risk reduction advocates. Another way of thinking about this is as a distinction between high-impact goals and high-variance strategies: once you’ve identified a high-impact goal, you can pursue it without using high-variance strategies. Startup X may have a crazy new business idea, but they probably shouldn't execute it in crazy new ways. Actually, their best bet is likely to be joining Y Combinator, getting a bunch of VC funding, and following Paul Graham's standard advice. Similarly, reducing x-risk is a crazy new idea for how to improve the world, but it's pretty plausible that we should pursue it in ways similar to those which other successful movements used. Here are some standard things that have historically been very helpful for changing the world:

- dedicated activists
- good research
- money
- public support
- political influence

My prior says that all of these things matter, and that most big wins will be due to direct effects on these things. The last two are the ones which we’re disproportionately lacking; I’m more optimistic about the latter for a variety of reasons.

AI policy is a particularly good place to have a large impact.

Here's a general argument: governments are very big levers, because of their scale and ability to apply coercion. A new law can be a black swan all by itself. When I think of really massive wins over the past half-century, I think about the eradication of smallpox and polio, the development of space technology, and the development of the internet. All of these relied on and were driven by governments. Then, of course, there are the massive declines in poverty across Asia in particular. It's difficult to assign credit for this, since it's so tied up with globalisation, but to the extent that any small group was responsible, it was Asian governments and the policies of Deng Xiaoping, Lee Kuan Yew, Rajiv Gandhi, etc.

You might agree that governments do important things, but think that influencing them is very difficult. Firstly, that's true for most black swans, so I don't think that should make policy work much less promising even from Ben's perspective. But secondly, from the outside view, our chances are pretty good. We're a movement comprising many very competent, clever and committed people. We've got the sort of backing that makes policymakers take people seriously: we're affiliated with leading universities, tech companies, and public figures. It's likely that a number of EAs at the best universities already have friends who will end up in top government positions. We

have enough money to do extensive lobbying, if that's judged a good idea. Also, we're correct, which usually helps. The main advantage we're missing is widespread popular support, but I don't model this as being crucial for issues where what's needed is targeted interventions which "pull the rope sideways". (We're also missing knowledge about what those interventions should be, but that makes policy research even more valuable).

Here's a more specific route to impact: in a few decades (assuming long timelines and slow takeoff) AIs that are less generally intelligent that humans will be causing political and economic shockwaves, whether that's via mass unemployment, enabling large-scale security breaches, designing more destructive weapons, psychological manipulation, or something even less predictable. At this point, governments will panic and AI policy advisors will have real influence. If competent and aligned people were the obvious choice for those positions, that'd be fantastic. If those people had spent several decades researching what interventions would be most valuable, that'd be even better.

This perspective is inspired by Milton Friedman, who argued that the way to create large-scale change is by nurturing ideas which will be seized upon in a crisis.

> Only a crisis - actual or perceived - produces real change. When that crisis occurs, the actions that are taken depend on the ideas that are lying around. That, I believe, is our basic function: to develop alternatives to existing policies, to keep them alive and available until the politically impossible becomes the possible.

The major influence of the Institute of Economic Affairs on Thatcher’s policies is an example of this strategy’s success. An advantage of this approach is that it can be implemented by clusterings of like-minded people collaborating with each other; for that reason, I'm not so worried about policy work cultivating the wrong mindset (I'd be more worried on this front if policy researchers were very widely spread out).

Another fairly specific route to impact: several major AI research labs would likely act on suggestions for coordinating to make AI safer, if we had any. Right now I don’t think we do, and so research into that could have a big multiplier. If a government ends up running a major AI lab (which seems pretty likely conditional on long timelines) then they may also end up following this advice, via the effect described in the paragraph above.

Underlying generators of this disagreement

More generally, Ben and I disagree on where the bottleneck to AI safety is. I think that finding a technical solution is probable, but that most solutions would still require careful oversight, which may or may not happen (maybe 50-50). Ben thinks that finding a technical solution is improbable, but that if it's found it'll probably be implemented well. I also have more credence on long timelines and slow takeoffs than he does. I think that these disagreements affect our views on the importance of influencing governments in particular.

We also have differing views on what the x-risk reduction community should look like. I favour a broader, more diverse community; Ben favours a narrower, more committed community. I don't want to discuss this extensively here, but I will point out that there are many people who are much better at working within a system than outside it - people who would do well in AI safety PhDs, but couldn't just teach themselves to do good research from scratch like Nate Soares did; brilliant yet absent-minded mathematicians; people who could run an excellent policy research group but not an

excellent startup. I think it's valuable for such people (amongst which I include myself), to have a "default" path to impact, even at the cost of reducing the pressure to be entrepreneurial or agenty. I think this is pretty undeniable when it comes to technical research, and cross-applies straightforwardly to policy research and advocacy.

Ben and I agree that going into policy is much more valuable if you're thinking very strategically and out of the "out of the box" box than if you're not. Given this mindset, there will probably turn out to be valuable non-standard things which you can do.

Do note that this essay is intrinsically skewed since I haven't portrayed Ben's arguments in full fidelity and have spent many more words arguing my side. Also note that, despite being skeptical about some of Ben's points, I think his overall view is important and interesting and more people should be thinking along similar lines.

Thanks to Anjali Gopal for comments on drafts.

The Rocket Alignment Problem

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

The following is a fictional dialogue building off of AI Alignment: Why It's Hard, and Where to Start.

(*Somewhere in a not-very-near neighboring world, where science took a very different course…*)

ALFONSO: Hello, Beth. I've noticed a lot of speculations lately about "spaceplanes" being used to attack cities, or possibly becoming infused with malevolent spirits that inhabit the celestial realms so that they turn on their own engineers.

I'm rather skeptical of these speculations. Indeed, I'm a bit skeptical that airplanes will be able to even rise as high as stratospheric weather balloons anytime in the next century. But I understand that your institute wants to address the potential problem of malevolent or dangerous spaceplanes, and that you think this is an important present-day cause.

BETH: That's… really not how we at the Mathematics of Intentional Rocketry Institute would phrase things.

The problem of malevolent celestial spirits is what all the news articles are focusing on, but we think the real problem is something entirely different. We're worried that there's a difficult, theoretically challenging problem which modern-day rocket punditry is mostly overlooking. We're worried that if you aim a rocket at where the Moon is in the sky, and press the launch button, the rocket may not actually end up at the Moon.

ALFONSO: I understand that it's very important to design fins that can stabilize a spaceplane's flight in heavy winds. That's important spaceplane safety research and someone needs to do it.

But if you were working on that sort of safety research, I'd expect you to be collaborating tightly with modern airplane engineers to test out your fin designs, to demonstrate that they are actually useful.

BETH: Aerodynamic designs are important features of any safe rocket, and we're quite glad that rocket scientists are working on these problems and taking safety seriously. That's not the sort of problem that we at MIRI focus on, though.

ALFONSO: What's the concern, then? Do you fear that spaceplanes may be developed by ill-intentioned people?

BETH: That's not the failure mode we're worried about right now. We're more worried that right now, *nobody* can tell you how to point your rocket's nose such that it goes to the moon, nor indeed *any* prespecified celestial destination. Whether Google or the US Government or North Korea is the one to launch the rocket won't make a pragmatic difference to the probability of a successful Moon landing from our

perspective, because right now *nobody knows how to aim any kind of rocket anywhere*.

ALFONSO: I'm not sure I understand.

BETH: We're worried that even if you aim a rocket at the Moon, such that the nose of the rocket is clearly lined up with the Moon in the sky, the rocket won't go to the Moon. We're not sure what a realistic path from the Earth to the moon looks like, but we suspect it might not be a very straight path, and it may not involve pointing the nose of the rocket at the moon at all. We think the most important thing to do next is to advance our understanding of rocket trajectories until we have a better, deeper understanding of what we've started calling the "rocket alignment problem". There are other safety problems, but this rocket alignment problem will probably take the most total time to work on, so it's the most urgent.

ALFONSO: Hmm, that sounds like a bold claim to me. Do you have a reason to think that there are invisible barriers between here and the moon that the spaceplane might hit? Are you saying that it might get very very windy between here and the moon, more so than on Earth? Both eventualities could be worth preparing for, I suppose, but neither seem likely.

BETH: We don't think it's particularly likely that there are invisible barriers, no. And we don't think it's going to be especially windy in the celestial reaches — quite the opposite, in fact. The problem is just that we don't yet know how to plot *any* trajectory that a vehicle could realistically take to get from Earth to the moon.

ALFONSO: Of course we can't plot an actual trajectory; wind and weather are too unpredictable. But your claim still seems too strong to me. Just aim the spaceplane at the moon, go up, and have the pilot adjust as necessary. Why wouldn't that work? Can you prove that a spaceplane aimed at the moon won't go there?

BETH: We don't think we can *prove* anything of that sort, no. Part of the problem is that realistic calculations are extremely hard to do in this area, after you take into account all the atmospheric friction and the movements of other celestial bodies and such. We've been trying to solve some drastically simplified problems in this area, on the order of assuming that there is no atmosphere and that all rockets move in perfectly straight lines. Even those unrealistic calculations strongly suggest that, in the much more complicated real world, just pointing your rocket's nose at the Moon also won't make your rocket end up at the Moon. I mean, the fact that the real world is more complicated doesn't exactly make it any *easier* to get to the Moon.

ALFONSO: Okay, let me take a look at this "understanding" work you say you're doing…

Huh. Based on what I've read about the math you're trying to do, I can't say I understand what it has to do with the Moon. Shouldn't helping spaceplane pilots exactly target the Moon involve looking through lunar telescopes and studying exactly what the Moon looks like, so that the spaceplane pilots can identify particular features of the landscape to land on?

BETH: We think our present stage of understanding is much too crude for a detailed Moon map to be our next research target. We haven't yet advanced to the point of targeting one crater or another for our landing. We can't target *anything* at this point. It's more along the lines of "figure out how to talk mathematically about curved rocket

trajectories, instead of rockets that move in straight lines". Not even realistically curved trajectories, right now, we're just trying to get past straight lines at all –

ALFONSO: But planes on Earth move in curved lines all the time, because the Earth itself is curved. It seems reasonable to expect that future spaceplanes will also have the capability to move in curved lines. If your worry is that spaceplanes will only move in straight lines and miss the Moon, and you want to advise rocket engineers to build rockets that move in curved lines, well, that doesn't seem to me like a great use of anyone's time.

BETH: You're trying to draw much too direct of a line between the math we're working on right now, and actual rocket designs that might exist in the future. It's *not* that current rocket ideas are almost right, and we just need to solve one or two more problems to make them work. The conceptual distance that separates anyone from solving the rocket alignment problem is *much greater* than that.

Right now everyone is *confused* about rocket trajectories, and we're trying to become *less confused*. That's what we need to do next, not run out and advise rocket engineers to build their rockets the way that our current math papers are talking about. Not until we stop being *confused* about extremely basic questions like why the Earth doesn't fall into the Sun.

ALFONSO: I don't think the Earth is going to collide with the Sun anytime soon. The Sun has been steadily circling the Earth for a long time now.

BETH: I'm not saying that our goal is to address the risk of the Earth falling into the Sun. What I'm trying to say is that if humanity's present knowledge can't answer questions like "Why doesn't the Earth fall into the Sun?" then we don't know very much about celestial mechanics and we won't be able to aim a rocket through the celestial reaches in a way that lands softly on the Moon.

As an example of work we're presently doing that's aimed at improving our understanding, there's what we call the " tiling positions " problem. The tiling positions problem is how to fire a cannonball from a cannon in such a way that the cannonball circumnavigates the earth over and over again, "tiling" its initial coordinates like repeating tiles on a tessellated floor –

ALFONSO: I read a little bit about your work on that topic. I have to say, it's hard for me to see what firing things from cannons has to do with getting to the Moon. Frankly, it sounds an awful lot like Good Old-Fashioned Space Travel, which everyone knows doesn't work. Maybe Jules Verne thought it was possible to travel around the earth by firing capsules out of cannons, but the modern study of high-altitude planes has completely abandoned the notion of firing things out of cannons. The fact that you go around talking about firing things out of cannons suggests to me that you haven't kept up with all the innovations in airplane design over the last century, and that your spaceplane designs will be completely unrealistic.

BETH: We know that rockets will not actually be fired out of cannons. We really, really know that. We're intimately familiar with the reasons why nothing fired out of a modern cannon is ever going to reach escape velocity. I've previously written several sequences of articles in which I describe why cannon-based space travel doesn't work.

ALFONSO: But your current work is all about firing something out a cannon in such a way that it circles the earth over and over. What could that have to do with any

realistic advice that you could give to a spaceplane pilot about how to travel to the Moon?

BETH: Again, you're trying to draw much too straight a line between the math we're doing right now, and direct advice to future rocket engineers.

We think that if we could find an angle and firing speed such that an ideal cannon, firing an ideal cannonball at that speed, on a perfectly spherical Earth with no atmosphere, would lead to that cannonball entering what we would call a "stable orbit" without hitting the ground, then… we might have understood something really fundamental and important about celestial mechanics.

Or maybe not! It's hard to know in advance which questions are important and which research avenues will pan out. All you can do is figure out the next tractable-looking problem that confuses you, and try to come up with a solution, and hope that you'll be less confused after that.

ALFONSO: You're talking about the cannonball hitting the ground as a problem, and how you want to avoid that and just have the cannonball keep going forever, right? But real spaceplanes aren't going to be aimed at the ground in the first place, and lots of regular airplanes manage to not hit the ground. It seems to me that this "being fired out of a cannon and hitting the ground" scenario that you're trying to avoid in this "tiling positions problem" of yours just isn't a failure mode that real spaceplane designers would need to worry about.

BETH: We are not worried about real rockets being fired out of cannons and hitting the ground. That is not why we're working on the tiling positions problem. In a way, you're being far too optimistic about how much of rocket alignment theory is already solved! We're not so close to understanding how to aim rockets that the kind of designs people are talking about now *would* work if only we solved a particular set of remaining difficulties like not firing the rocket into the ground. You need to go more meta on understanding the kind of progress we're trying to make.

We're working on the tiling positions problem because we think that being able to fire a cannonball at a certain instantaneous velocity such that it enters a stable orbit… is the sort of problem that somebody who could really actually launch a rocket through space and have it move in a particular curve that really actually ended with softly landing on the Moon would be able to solve *easily*. So the fact that we can't solve it is alarming. If we can figure out how to solve this much simpler, much more crisply stated "tiling positions problem" with imaginary cannonballs on a perfectly spherical earth with no atmosphere, which is a lot easier to analyze than a Moon launch, we might thereby take one more incremental step towards eventually becoming the sort of people who could plot out a Moon launch.

ALFONSO: If you don't think that Jules-Verne-style space cannons are the wave of the future, I don't understand why you keep talking about cannons in particular.

BETH: Because there's a lot of sophisticated mathematical machinery already developed for aiming cannons. People have been aiming cannons and plotting cannonball trajectories since the sixteenth century. We can take advantage of that existing mathematics to say exactly how, if we fired an ideal cannonball in a certain direction, it would plow into the ground. If we tried talking about rockets with realistically varying acceleration, we can't even manage to prove that a rocket like that *won't* travel around the Earth in a perfect square, because with all that realistically varying acceleration and realistic air friction it's impossible to make any

sort of definite statement one way or another. Our present understanding isn't up to it.

ALFONSO: Okay, another question in the same vein. Why is MIRI sponsoring work on adding up lots of tiny vectors? I don't even see what that has to do with rockets in the first place; it seems like this weird side problem in abstract math.

BETH: It's more like… at several points in our investigation so far, we've run into the problem of going from a function about time-varying accelerations to a function about time-varying positions. We kept running into this problem as a blocking point in our math, in several places, so we branched off and started trying to analyze it explicitly. Since it's about the pure mathematics of points that don't move in discrete intervals, we call it the "logical undiscreteness" problem. Some of the ways of investigating this problem involve trying to add up lots of tiny, varying vectors to get a big vector. Then we talk about how that sum seems to change more and more slowly, approaching a limit, as the vectors get tinier and tinier and we add up more and more of them… or at least that's one avenue of approach.

ALFONSO: I just find it hard to imagine people in future spaceplane rockets staring out their viewports and going, "Oh, no, we don't have tiny enough vectors with which to correct our course! If only there was some way of adding up even more vectors that are even smaller!" I'd expect future calculating machines to do a pretty good job of that already.

BETH: Again, you're trying to draw much too straight a line between the work we're doing now, and the implications for future rocket designs. It's not like we think a rocket design will almost work, but the pilot won't be able to add up lots of tiny vectors fast enough, so we just need a faster algorithm and then the rocket will get to the Moon. This is foundational mathematical work that we think might play a role in multiple basic concepts for understanding celestial trajectories. When we try to plot out a trajectory that goes all the way to a soft landing on a moving Moon, we feel confused and blocked. We think part of the confusion comes from not being able to go from acceleration functions to position functions, so we're trying to resolve our confusion.

ALFONSO: This sounds suspiciously like a philosophy-of-mathematics problem, and I don't think that it's possible to progress on spaceplane design by doing philosophical research. The field of philosophy is a stagnant quagmire. Some philosophers still believe that going to the moon is impossible; they say that the celestial plane is fundamentally separate from the earthly plane and therefore inaccessible, which is clearly silly. Spaceplane design is an engineering problem, and progress will be made by engineers.

BETH: I agree that rocket design will be carried out by engineers rather than philosophers. I also share some of your frustration with philosophy in general. For that reason, we stick to well-defined mathematical questions that are likely to have actual answers, such as questions about how to fire a cannonball on a perfectly spherical planet with no atmosphere such that it winds up in a stable orbit.

This often requires developing new mathematical frameworks. For example, in the case of the logical undiscreteness problem, we're developing methods for translating between time-varying accelerations and time-varying positions. You can call the development of new mathematical frameworks "philosophical" if you'd like — but if you do, remember that it's a very different kind of philosophy than the "speculate

about the heavenly and earthly planes" sort, and that we're always pushing to develop new mathematical frameworks or tools.

ALFONSO: So from the perspective of the public good, what's a good thing that might happen if you solved this logical undiscreteness problem?

BETH: Mainly, we'd be less confused and our research wouldn't be blocked and humanity could actually land on the Moon someday. To try and make it more concrete – though it's hard to do that without actually knowing the concrete solution – we might be able to talk about incrementally more realistic rocket trajectories, because our mathematics would no longer break down as soon as we stopped assuming that rockets moved in straight lines. Our math would be able to talk about exact curves, instead of a series of straight lines that approximate the curve.

ALFONSO: An exact curve that a rocket follows? This gets me into the main problem I have with your project in general. I just don't believe that any future rocket design will be the sort of thing that can be analyzed with absolute, perfect precision so that you can get the rocket to the Moon based on an absolutely plotted trajectory with no need to steer. That seems to me like a bunch of mathematicians who have no clue how things work in the real world, wanting everything to be perfectly calculated. Look at the way Venus moves in the sky; usually it travels in one direction, but sometimes it goes retrograde in the other direction. We'll just have to steer as we go.

BETH: That's not what I meant by talking about exact curves… Look, even if we can invent logical undiscreteness, I agree that it's futile to try to predict, in advance, the precise trajectories of all of the winds that will strike a rocket on its way off the ground. Though I'll mention parenthetically that things might actually become calmer and easier to predict, once a rocket gets sufficiently high up –

ALFONSO: Why?

BETH: Let's just leave that aside for now, since we both agree that rocket positions are hard to predict exactly during the atmospheric part of the trajectory, due to winds and such. And yes, if you can't exactly predict the initial trajectory, you can't exactly predict the later trajectory. So, indeed, the proposal is definitely not to have a rocket design so perfect that you can fire it at exactly the right angle and then walk away without the pilot doing any further steering. The point of doing rocket math isn't that you want to predict the rocket's exact position at every microsecond, in advance.

ALFONSO: Then why obsess over pure math that's too simple to describe the rich, complicated real universe where sometimes it rains?

BETH: It's true that a real rocket isn't a simple equation on a board. It's true that there are all sorts of aspects of a real rocket's shape and internal plumbing that aren't going to have a mathematically compact characterization. What MIRI is doing isn't the right degree of mathematization for all rocket engineers for all time; it's the mathematics for us to be using right now (or so we hope).

To build up the field's understanding incrementally, we need to talk about ideas whose consequences can be pinpointed precisely enough that people can analyze scenarios in a shared framework. We need enough precision that someone can say, "I think in scenario X, design Y does Z", and someone else can say, "No, in scenario X, Y actually does W", and the first person responds, "Darn, you're right. Well, is there some way to change Y so that it would do Z?"

If you try to make things realistically complicated at this stage of research, all you're left with is verbal fantasies. When we try to talk to someone with an enormous flowchart of all the gears and steering rudders they think should go into a rocket design, and we try to explain why a rocket pointed at the Moon doesn't necessarily end up at the Moon, they just reply, "Oh, my rocket won't do *that*." Their ideas have enough vagueness and flex and underspecification that they've achieved the safety of nobody being able to prove to them that they're wrong. It's impossible to incrementally build up a body of collective knowledge that way.

The goal is to start building up a library of tools and ideas we can use to discuss trajectories formally. Some of the key tools for formalizing and analyzing *intuitively* plausible-seeming trajectories haven't yet been expressed using math, and we can live with that for now. We still try to find ways to represent the key ideas in mathematically crisp ways whenever we can. That's not because math is so neat or so prestigious; it's part of an ongoing project to have arguments about rocketry that go beyond "Does not!" vs. "Does so!"

ALFONSO: I still get the impression that you're reaching for the warm, comforting blanket of mathematical reassurance in a realm where mathematical reassurance doesn't apply. We can't obtain a mathematical certainty of our spaceplanes being absolutely sure to reach the Moon with nothing going wrong. That being the case, there's no point in trying to pretend that we can use mathematics to get absolute guarantees about spaceplanes.

BETH: Trust me, I am not going to feel "reassured" about rocketry no matter what math MIRI comes up with. But, yes, of course you can't obtain a mathematical assurance of any physical proposition, nor assign probability 1 to any empirical statement.

ALFONSO: Yet you talk about proving theorems – proving that a cannonball will go in circles around the earth indefinitely, for example.

BETH: Proving a theorem about a rocket's trajectory won't ever let us feel comfortingly certain about where the rocket is actually going to end up. But if you can prove a theorem which says that your rocket would go to the Moon if it launched in a perfect vacuum, maybe you can attach some steering jets to the rocket and then have it actually go to the Moon in real life. Not with 100% probability, but with probability greater than zero.

The point of our work isn't to take current ideas about rocket aiming from a 99% probability of success to a 100% chance of success. It's to get past an approximately 0% chance of success, which is where we are now.

ALFONSO: Zero percent?!

BETH: Modulo Cromwell's Rule, yes, zero percent. If you point a rocket's nose at the Moon and launch it, it does not go to the Moon.

ALFONSO: I don't think future spaceplane engineers will actually be that silly, if direct Moon-aiming isn't a method that works. They'll lead the Moon's current motion in the sky, and aim at the part of the sky where Moon will appear on the day the spaceplane is a Moon's distance away. I'm a bit worried that you've been talking about this problem so long without considering such an obvious idea.

BETH: We considered that idea very early on, and we're pretty sure that it still doesn't get us to the Moon.

ALFONSO: What if I add steering fins so that the rocket moves in a more curved trajectory? Can you prove that no version of that class of rocket designs will go to the Moon, no matter what I try?

BETH: Can you sketch out the trajectory that you think your rocket will follow?

ALFONSO: It goes from the Earth to the Moon.

BETH: In a bit more detail, maybe?

ALFONSO: No, because in the real world there are always variable wind speeds, we don't have infinite fuel, and our spaceplanes don't move in perfectly straight lines.

BETH: Can you sketch out a trajectory that you think a simplified version of your rocket will follow, so we can examine the assumptions your idea requires?

ALFONSO: I just don't believe in the general methodology you're proposing for spaceplane designs. We'll put on some steering fins, turn the wheel as we go, and keep the Moon in our viewports. If we're off course, we'll steer back.

BETH: … We're actually a bit concerned that standard steering fins may stop working once the rocket gets high enough, so you won't actually find yourself able to correct course by much once you're in the celestial reaches – like, if you're already on a good course, you can correct it, but if you screwed up, you won't just be able to turn around like you could turn around an airplane –

ALFONSO: Why not?

BETH: We can go into that topic too; but even given a simplified model of a rocket that you *could* steer, a walkthrough of the steps along the path that simplified rocket would take to the Moon would be an important step in moving this discussion forward. Celestial rocketry is a domain that we expect to be unusually difficult – even compared to building rockets on Earth, which is already a famously hard problem because they usually just explode. It's not that everything has to be neat and mathematical. But the overall difficulty is such that, in a proposal like "lead the moon in the sky," if the core ideas don't have a certain amount of solidity about them, it would be equivalent to firing your rocket randomly into the void.

If it feels like you don't know for sure whether your idea works, but that it might work; if your idea has many plausible-sounding elements, and to you it feels like nobody has been able to *convincingly* explain to you how it would fail; then, in real life, that proposal has a roughly 0% chance of steering a rocket to the Moon.

If it seems like an idea is extremely solid and clearly well-understood, if it feels like this proposal should definitely take a rocket to the Moon without fail in good conditions, then maybe under the best-case conditions we should assign an 85% subjective credence in success, or something in that vicinity.

ALFONSO: So uncertainty automatically means failure? This is starting to sound a bit paranoid, honestly.

BETH: The idea I'm trying to communicate is something along the lines of, "If you can reason rigorously about why a rocket should definitely work in principle, it might work in real life, but if you have anything less than that, then it definitely won't work in real life."

I'm not asking you to give me an absolute mathematical proof of empirical success. I'm asking you to give me something more like a sketch for how a simplified version of your rocket could move, that's sufficiently determined in its meaning that you can't just come back and say "Oh, I didn't mean *that*" every time someone tries to figure out what it actually does or pinpoint a failure mode.

This isn't an unreasonable demand that I'm imposing to make it impossible for any ideas to pass my filters. It's the primary bar all of us have to pass to contribute to collective progress in this field. And a rocket design which can't even pass that conceptual bar has roughly a 0% chance of landing softly on the Moon.

Public Positions and Private Guts

In this post I lay out a model of beliefs and communication that identify two types of things we might think of as 'beliefs,' how they are communicated between people, how they are communicated within people, and what this might imply about intellectual progress in some important fields. As background, Terence Tao has a blog post describing three stages of mathematics: pre-rigorous, rigorous, and post-rigorous. It's only about two pages; the rest of this post will assume you've read it. Ben Pace has a blog post describing how to discuss the models generating an output, rather than the output itself, which is also short and is related, but has important distinctions from the model outlined here.

[Note: the concept for this post comes from a talk given by Anna Salamon, and I sometimes instruct for CFAR, but the presentation in this post should be taken to only represent my views.]

> If a man will begin with certainties, he shall end with doubts, but if he will be content to begin with doubts he shall end in certainties. -- Francis Bacon

FORMAL COMMUNICATION

Probably the dominant model of conversations among philosophers today is Robert Stalnaker's. (Here's an introduction.) A conversation has a defined set of interlocutors, and some shared context, and speech acts add statements to the context, typically by asserting a new fact.

I'm not an expert in contemporary philosophy, and so from here on out this is my extension of this view that I'll refer to as 'formal.' Perhaps this extension is entirely precedented, or perhaps it's controversial. My view focuses on situations where logical omniscience is not assumed, and thus simply pointing out the conclusion that arises from combining facts can count as such an assertion. Proper speech considers this and takes inferential distance into account; my speech acts should be derivable from our shared context or an unsurprising jump from them. Both new logical facts and environmental facts count as adding information to the shared context. That I am currently wearing brown socks while writing this part of the post is not something you could derive from our shared context, but is nevertheless 'unsurprising.'

It's easy to see how a mathematical proof might fit into this framework. We begin with some axioms and suppositions, and then we compute conclusions that follow from those premises, and eventually we end up at the theorem that was to be proved.

If I make a speech act that's too far of a stretch--either because it disagrees with something in the context (or your personal experience), or is just not easily derivable from the common context--then the audience should 'beep' and I should back up and justify the speech act. A step in the proof that doesn't obviously follow means I need to expand the proof to make it clear how I got from A to B, or how a pair of statements that appear contradictory is in fact not contradictory. ("Ah, by X I meant the restricted subset X', such that this counterexample is excluded; my mistake.")

This style of conversation seems to be minimizing surprise on the low level; from moment to moment, actions are being taken in a way that views justification and

validation by independent sources as core constraints. What is this good for? Interestingly, the careful avoidance of surprises on the low level permits surprises on the high level, as a conclusion reached by airtight logic can be as trustworthy as the premises of that logic, regardless of how bizarre the conclusion seems. A plan fleshed out with enough detail that it can be independently reconstructed by many different people is a plan that can scale to a large organization. The body of scientific knowledge is communicated mostly this way; Nullius in verba requires this sort of careful communication because it bans the leaps one might otherwise make.

PUBLIC POSITIONS

One way to model communication is a function that takes objects of a certain type and tries to recreate them in another place. A telephone takes sound waves and attempts to recreate them elsewhere, whereas an instant messenger takes text strings and attempts to recreate them elsewhere. So conjugate to the communication methodology is 'the thing that can be communicated by this methodology'. For example, you can also play music over the telephone, which is much harder to do over instant messenger (tho this example perhaps betrays my age).

I'm going to define 'public positions' as the sort of beliefs that are amenable to communication through 'formal communication' (this style where you construct conclusions out of a chain of simple additions to the pre-existing context). The 'public' bit emphasizes that they're optimized for justification or presentation; many things I believe don't count as public positions because I can't reach them through this sort of formal communication. For example, I find the smell of oranges highly unpleasant; I can communicate that fact about my preferences through formal communication but can't communicate the preference itself through formal communication. The 'positions' bit emphasizes that they are *defensible* and *legible*; you can 'know where I stand' on a particular topic.

PRIVATE GUTS

I'm going to call a different sort of belief one's 'private guts.' By 'guts,' I'm pointing towards the historical causes of a belief (like the particular bit of my biochemistry that causes me to dislike the smell of oranges), or to the sense of a 'gut feeling.' By private, I'm pointing towards the fact that this is often opaque or not shaped like something that's communicable, rather than something deliberately hidden. If you're familiar with Gendlin's Focusing, 'felt senses' are an example of private guts.

What are private guts good for? As far as I can tell, lizards probably don't have public positions, but they probably do have private guts. That suggests those guts are good for predicting things about the world and achieving desirable world states, as well as being one of the channels by which the desirability of world states is communicated inside a mind. It seems related to many sorts of 'embodied knowledge', like how to walk, which is not understood from first principles or in an abstract way, or habits, like adjective order in English. A neural network that 'knows' how to classify images of cats, but doesn't know how it knows (or is 'uninterpretable'), seems like an example of this. "Why is this image a cat?" -> "Well, because when you do lots of multiplication and addition and nonlinear transforms on pixel intensities, it ends up having a higher cat-number than dog-number." This seems similar to gut senses that are difficult to articulate; "why do you think the election will go this way instead of that way?" ->

“Well, because when you do lots of multiplication and addition and nonlinear transforms on environmental facts, it ends up having a higher A-number than B-number.” Private guts also seem to capture a category of amorphous visions; a startup can rarely write a formal proof that their project will succeed (generally, if they could, the company would already exist). The postrigorous mathematician’s hunch falls into this category, which I’ll elaborate on later.

There are now two sorts of interesting communication to talk about: the process that coheres public positions and private guts within a single individual, and the process that communicates private guts across individuals.

COHERENCE, FOCUSING, AND SCIENCE

Much of CFAR’s focus, and that of the rationality project in general, has involved taking people who are extremely sophisticated at formal communication and developing their public positions, and getting them to notice and listen to their private guts. An example, originally from Julia Galef, is the ‘agenty duck.’ Imagine a duck whose head points in one direction (“I want to get a PhD!”) and whose feet are pointed in another (mysteriously, this duck never wants to work on their dissertation). Many responses to this sort of intrapersonal conflict seem maladaptive; much better for the duck to have head and feet pointed in the same direction, regardless of which direction that is. An individual running a coherence process that integrates the knowledge of the ‘head’ and ‘feet’, or the public positions and the private guts, will end up more knowledgeable and functional than an individual that ignores one to focus on the other.

Discovering the right coherence process is an ongoing project, and even if I knew it as a public position it would be too long for this post. So I will merely leave some pointers and move on. First, the private guts seem highly trainable by experience, especially through carefully graduated exposure. Second, Focusing and related techniques (like Internal Double Crux) seem quite effective at searching through the space of articulable / understandable sentences or concepts in order to find those that resonate with the private guts, drawing forth articulation from the inarticulate.

It’s also worth emphasizing the way in which science depends on such a coherence process. The ‘scientific method’ can be viewed in this fashion: hypotheses can be wildly constructed through any method, because hypotheses are simply proposals rather than truth-statements; only hypotheses that survive the filter of contact with reality through experimentation graduate to full facts, at which point their origin is irrelevant, be it induction, a lucky guess, or the unconscious mind processing something in a dream.

Similarly for mathematicians, according to Tao. The transition from pre-rigorous mathematics to rigorous mathematics corresponds to being able to see formal communication and public positions as types, and learning to trust them over persuasion and opinions. The transition from rigorous mathematics to post-rigorous mathematics corresponds to having trained one’s private guts such that they line up with the underlying mathematical reality well enough that they generate fruitful hypotheses.

Consider automatic theorem provers. One variety begins with a set of axioms, including the negated conclusion, and then gradually expands outwards, seeking to find a contradiction (and thus prove that the desired conclusion follows from the other

axioms). Every step of the way proceeds according to the formal communication style, and every proposition in the proof state can be justified through tracing the history of combinations of propositions that led from the initial axioms to that proposition. But the process is unguided, reliant on the swiftness of computer logic to handle the massive explosion of propositions, almost all of which will be irrelevant to the final proof. The human mathematician instead has some amorphous sense of what the proof will look like, sketching a leaky argument that is not correct in the details, but which is correctable. Something interesting is going on in the process that generates correctable arguments, perhaps even more interesting than what's going on in the processes that trivially generate correct arguments by generating all possible arguments and then filtering.

STARTUPS, DOUBLE CRUX, AND CIRCLING

Somehow, people are sometimes able to link up their private guts with each other. This is considerably more fraught than linking up public positions; positions are of a type that is optimized for verifiability and reconstruction, whereas internal experiences, in general, are not. Even if we're eating the same cake, how would we even check that our internal experience of eating the cake is similar? What about something simpler, like seeing the same color?

While the abstraction of formal conversation is fairly simple, it's still obvious that there are many skills related to correct argumentation. Similarly, there seems to be a whole family of skills related to syncing up private guts, and rather than teaching those skills, this section will again by a pointer to where those skills could be learned or trained. Learning how to reproduce music is related to learning how to participate in jam sessions, but the latter is a much closer fit to this sort of communication.

The experience of startups is that small teams are best, primarily because of the costs of coordinative communication. Startups are often chasing an amorphous, rapidly changing target; a team that's able to quickly orient in the same direction and move together, or trust in the guts of each other rather than requiring elaborate proofs, will often perform better.

While Double Crux can generate a crisp tree of logical deductions from factual disagreements, it often instead exposes conflicting intuitions or interpretations. While formal communication involves a speaker optimizing over speech acts to jointly minimize surprise and maximize distance towards their goal, double crux instead involves both parties in the optimization, and often causes them to *seek* surprises. A crux is something that would change my mind, and I expose my cruxes in case you disagree with them, seeking to change *my* mind as quickly as possible.

Cruxes also respect the historical causes of beliefs; when I say "my crux for X is Y," I am not saying that Y should cause you to believe X, only that not-Y would cause me to believe not-X. This weaker filter means many more statements are permissible, and my specific epistemic state can be addressed, rather than playing a minimax game by which all possible interlocutors would be pinned down by the truth. In Stalnakerian language, rather than needing to emit statements that are understandable and justifiable by the common context, I only need the weaker restriction that those statements are understandable in the common context and justifiable in my private context.

Circling is also beyond the scope of this post, except as a pointer. It seems relevant as a potential avenue for deliberate practice in understanding and connecting to the subjective experience of others in a way that perhaps facilitates this sort of conversation.

CONCLUSION

As mentioned, this post is seeking to set out a typology, and perhaps crystallize some concepts. But why think these concepts are useful?

Primarily, because this seems to be related to the way in which rationalists differ from other communities with similar interests, or from their prior selves before coming a rationalist, in a way that seems related to the difference between postrigorous mathematicians and rigorous mathematicians. Secondarily, because many contemporary issues of great practical importance require correctly guessing matters that are not settled. Financial examples are easy (“If I buy bitcoin now, will it be worth more when I sell it by more than a normal rate of economic return?”), but longevity interventions have a similar problem (“knowing whether or not this works for humans will take a human lifetime to figure out, but by that point it might be too late for me. Should I do it now?”), and it seems nearly impossible to reason correctly about existential risks without reliance on private guts (and thus on methods to tune and communicate those guts).

Good Samaritans in experiments

Consider 2 people. Both are seminary students who are taking part in an experiment ostensibly to consider different types of religiosity. One is asked to prepare a short talk on the Good Samaritan, the other on potential future careers for seminar graduates.

They are both told to go to another room to record their talk. The one who is to be giving a talk on the Good Samaritan is told that he is late and needs to hurry. The other participant is told that he has a time to spare.

If they, separately, come across someone who appears to be in respiratory distress, which do you think is more likely to stop and help?

Does being in a hurry determine whether someone helps?

Does reading the Good Samaritan?

Which is a bigger effect?

I was recently told about an experiment which showed that seminary students who had just prepared to give a talk about the Good Samaritan were no more likely to help someone in need than those who had been preparing a talk about an unrelated topic.

This seemed unexpected to me – people who had just been reading and thinking about a story which was told specifically by the leader of their faith to instruct them to help other people were no more likely to help than the control? I know humanity is crazy but seemed like a new level of crazy which I wouldn't have predicted.

So I thought I'd check out the study and – Aaaaaaaaaaaaaaaaah!!!

I know getting overly upset about bad experiments (especially those from before the replication crisis) is probably bad for my health but still – Aaaaaaaaaaaaaaaaaaaaaaah!!!

I don't want to be too harsh on the authors as this probably isn't the worst culprit you'll see but – Aaaaaaaaaaaaaah!!!

The paper has 1811 citations listed on google scholar – Aaaaaaaaaaaaaah!!!

I'm tempted to pretend that this post has some purpose other than just as a release of my frustration but that would be dishonest. Please consider this post a form of therapy for me. The working title for this post was "Screaming into the void" - consider yourself warned.

(If you want a more coherent discussion of common misuse of statistics in research papers I highly recommend putanumonit's defense against the dark arts series)

The Experiment

Ok, so the basic premise of the experiment seems to be sound. We want to know what inputs cause people to be more or less likely to help others:

1. Planning a talk on the Good Samaritan (GS)

2. Being in a hurry

3. Type of religiosity (Religion as quest, means or end)

The setup is to give people a questionnaire to determine their type of religiosity. Then give them some time to plan a short talk (3-5 mins) on GS or an unrelated topic. They are then asked to go to another room to give the talk (with 3 degrees of urgency – low, medium and high).

Contrary to the example given in the introduction, the level of hurriedness doesn't depend on which topic the individual has prepared – there are 6 conditions people are put in: GS low, medium and high urgency and control low, medium and high urgency.

On the way to the other room, you arrange for them to come across someone slumped in a doorway, with an apparent respiratory condition.

You monitor the subjects' responses and analyse the results.

My first question was whether they would adjust their p-value requirement for the 5 variables they were testing but no, it turns out that $p<0.05$ was deemed adequate for significance. Ok, could be worse I guess. More on this later.

The second place where doubts started to creep in were the rankings of responses:

0 = failed to notice the victim as possibly in need at all;

1 = perceived the victim as possibly in need but did not offer aid;

2 = did not stop but helped indirectly (e.g., by telling Steiner's assistant about the victim);

3 = stopped and asked if victim needed help;

4 = after stopping, insisted on taking the victim inside and then left him.

5 = after stopping, refused to leave the victim (after 3-5 minutes) and/or insisted on taking him somewhere outside experimental context (e.g., for coffee or to the infirmary).

It seems to me that there are two different things being measured here:

How likely are they to notice someone in distress

If they notice, how likely are they to help

However, these two factors have been put on the same scale. Surely it would be far more informative to separate them – does being in a hurry make you less likely to help someone you see who is in need or does it just make you less observant?

But this is relatively minor and the victim doesn't really mind why you're not helping, just whether you help or not – there could still be some useful results here. From the conclusion, it looks like not noticing the victim at all was rare so this failure mode was maybe not too big a deal.

Drawing the wrong conclusions

So, on to the conclusions from the report:

1. People who had been preparing GS talks were no more likely to help

2. People who were in a hurry were less likely to help

3. Religiosity type had an effect on the type of help given

It was the first which I was most interested in so I looked at the actual results.

53% of the people who had been preparing GS talks offered some kind of help (10/19).
29% of the people preparing non-GS talks offered some kind of help (6/21).

Wait, surely that means people who prepared a GS talk were 1.8x more likely to help than those with an alternative topic? Oh no, says the report. The difference was not significant at the $p<0.05$ level. Therefore, there is no effect. This isn't specifically stated in that way but "lack of significant effect" is immediately followed by "lack of effect":

> Although the degree to which a person was in a hurry had a clearly significant effect on his likelihood of offering the victim help, whether he was going to give a sermon on the parable or on possible vocational roles of ministers did not. This lack of effect of sermon topic raises certain difficulties for an explanation of helping behavior involving helping norms and their salience.

The paper goes some way to redeeming itself by stating:

> The results were in the direction suggested by the norm salience hypothesis, but they were not significant. The most accurate conclusion seems to be that salience of helping norms is a less strong determinant of helping behavior in the present situation than many, including the present authors, would expect.

It then undoes the good work in the next sentence:

> Thinking about the Good Samaritan did not increase helping behaviour

Part of me wants to be happy that they at least included a fairly accurate description of the evidence but the repeated stating of the incorrect conclusion throughout the report can only lead readers to the wrong conclusion.

At one point, the paper seems to go even further and claims that the fact that we can't reject the null hypothesis is confirmation of the null hypothesis:

> The prediction involved in the first hypothesis concerning the message content was based on the parable. The parable itself seemed to suggest that thinking pious thoughts would not increase helping. Another and conflicting prediction

> might be produced by a norm salience theory. Thinking about the parable should make norms for helping salient and therefore produce more helping. The data, as hypothesized, are more congruent with the prediction drawn from the parable. A person going to speak on the parable of the Good Samaritan is not significantly more likely to stop to help a person by the side of the road than is a person going to talk about possible occupations for seminary graduates.
>
> Since both situational hypotheses are confirmed…

Aaaaaaaaaaaaaaaaaaaaaaaaaaah!!!

Somehow, the paper manages to make the “pious thoughts are ineffective” hypothesis into the null hypothesis and the “norm salience” hypothesis into the alternative hypothesis. Then, when the results are not significant to reject the null hypothesis this is treated as confirmation that the null hypothesis is true. This is the equivalent of accepting $p<0.95$ as evidence for the “pious thoughts are ineffective” hypothesis.

(Aside: I’m no theologian but I’m not really sure that “pious thoughts are ineffective” is really what the parable implies. Jesus often used the religious leaders as the bad guys in his parables so he may just be repeating that point)

Effect Size, Significance and Experimental Power

I think the issue here is confusion between effect size and significance.

The effect size is actually pretty good (80% increase in helping). In fact, in the condition that the GS participants weren’t rushing they averaged an impressive score of 3.8 (compared to 1.667 for the equivalent non-GS participants).

The fact that this doesn’t rise to significance has little to do with effect size and everything to do with experimental power.

The sample size was 40. There were 6 categories relating to the first 2 hypotheses (3 hurry conditions x 2 message conditions). If for each of the 3 religiosity type conditions a participant was just rated as “high” or “low” then this is 8 categories. That makes a total of 48 possible categorisations for each subject to cover the 3 hypotheses. We’ve managed to get more potential categorisations of each subject than we have subjects.

Aaaaaaaaaaaaaaah!!!

(Actually, this may not be irretrievable in and of itself – it just threw up a big red flag for me. If all the other parts of the experiment were on the money this could just be efficiently testing as many different hypotheses as possible given the limited data points available. The real problem is that if we adjust for multiple variable testing then the required p-value for significance goes down and power goes down with it.)

In addition to sample size, experimental power depends on variation in the dependant variable due to other sources (I’m happy to accept that they had low measurement error). My best guess is that there is significant variation due to other sources

although I don’t have the data to show this. A number of personality traits had been investigated previously (Machiavellianism, authoritarianism, social desirability, alienation, and social responsibility) and found not to significantly correlate with helping behaviour, so my expectation would be that finding a true effect is difficult and unexplained variation in helping is large.

If experimental power is low, in order to find significant results, the effect size must be large.

As the effect size of reading GS was below the effect size required, the result is not statistically significant.

If an effect size of increasing helping by 80% is not significant, you really should have known before the experiment that you didn’t have enough power.

Further reducing experimental power

If you thought that N=40 was questionable, wait until you see what comes next. The paper goes on to see if the input variables correlate with the amount of help given when help was given. Only 16 people gave any help so suddenly N=16.

Aaaaaaaaaaaaaaaaaaaaaah!!!

This seems like bad news for finding significance but we suddenly do have a significant effect. It turns out that scoring higher on seeing religion as a quest makes you likely to offer less help than if your score lower on this metric. This is contrary to the experimenters’ expectations.

After performing some extra calculations, the experimenters conclude that this is because those who scored lower on this metric were likely to offer over-the-top assistance and score a 5 which skewed the results.

Allow me to offer an alternative explanation.

The paper has so far calculated 18 different p-values (3 from ANOVA of message x hurry, 10 from linear regression of full data (5 x help vs no help, 5 x scoring system) and 5 from linear regression of only helpful participants). There were actually another 10 p-values calculated in their stepwise multiple regression analysis but these seem to have been ignored so I’ll gloss over that.

Now for each p-value which you calculate you have a 5% chance of finding a spurious result. I’ll take off the 3 p-value calculations which yielded true effects and say 15 opportunities to get a spurious p-value.

$0.95 ^ 15 = 0.46$

At this point, you are more likely to have achieved a spurious p-value than not from all the calculated p-values. Some of the p-values calculated are related so that may change the exact value but the probability of a spurious result is uncomfortably high.

Remember that an increase in helping of 80% didn’t achieve significance when N=40. The effect size must be truly huge in order to achieve significance with N=16 (The actual effect size isn’t given in the report).

Because their prior for this effect being true is fairly low (it's huge and in the opposite direction to expectation) it would be reasonable to say that the p-value is probably spurious in the report with a note that this might be worth investigating further in the future.

Instead, the report ends up with a weird conclusion that low religion-as-a-quest scoring people are more likely to offer over-the-top help. The fact that they achieve an additional significant p-value when the introduce a new categorisation system (over-the-top help vs reasonable help) doesn't add much to the likelihood of their conclusion – it just shows that they are able to look at their data and see a pattern.

Introducing new variables

At this point, another input variable is introduced. The original 3 types of religiosity were made up of scores from 6 different scales which were weighted to create the 3 types. Suddenly one of the 6 original scales is grabbed out (doctrinal orthodoxy) and this correlates even more strongly with giving over-the-top help ($p<0.01$).

Aaaaaaaaaaaaaaaaaaah!!!

Introducing a new categorisation (over-the-top help) and a new variable (doctrinal orthodoxy) to try to explain a (probably) spurious p-value from multiple hypothesis testing is NOT a good idea.

We now have 4 different potential categorisations and 11 variables (the original 5 plus the 6 newly introduced scales). This makes 44 different potential p-values to calculate even before we consider the different types of tests that the authors might try (simple linear regression, ANOVA, stepwise multiple linear regression). I don't think they calculated all of these 44+ p-values but rather looked at the data and decided which ones looked promising.

$0.99 ^ 44 = 0.64$

So now, even in the best case, a $p<0.01$ would happen in more than a third of similar experiments just by coincidence.

I don't think that the effect described is impossible but I think the failure to adjust for multiple variables is a much more likely explanation.

Conclusion

So in conclusion, against all expectation, reading and preparing a talk on a parable given by the leader of your religion on how we should help people who are in need does, in fact, increase the likelihood that you will, in the next 5 minutes, help someone who is in need.

The fact that being in a hurry is a larger effect is the truly interesting finding here but I think not a huge surprise.

This is why I asked the question the way I did in the introduction – I didn't get the chance to guess this blind and I'm not sure which way I would have voted if I had.

I'm confident that I wouldn't have predicted quite such a big drop of help between GS low hurry and GS high hurry so I'll have to update accordingly (average score 3.8 down to score 1).

One final thing:

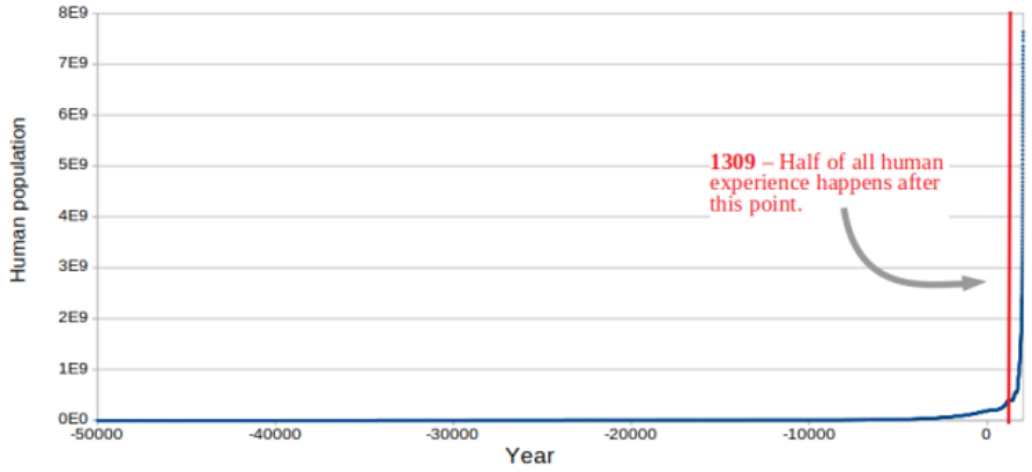
Aaaaaaaaaaaaaaaaaaaaaaaah!!!

The funnel of human experience

This is a linkpost for https://eukaryotewritesblog.com/2018/10/09/the-funnel-of-human-experience/

[EDIT: Previous version of this post had a major error. Thanks for jeff8765 for pinpointing the error and esrogs in the Eukaryote Writes Blog comments for bringing it to my attention as well. This has been fixed. Also, I wrote FHI when I meant FLI.]

The graph of the human population over time is also a map of human experience. Think of each year as being "amount of human lived experience that happened this year." On the left, we see the approximate dawn of the modern human species in 50,000 BC. On the right, the population exploding in the present day.



It turns out that if you add up all these years, 50% of human experience has happened after 1309 AD. 15% of all experience has been experienced by people who are alive right now.

I call this "the funnel of human experience" - the fact that because of a tiny initial population blossoming out into a huge modern population, more of human experience has happened recently than time would suggest.

50,000 years is a long time, but 8,000,000,000 people is a lot of people.

If you want to expand on this, you can start doing some Fermi estimates. We as a species have spent...

- 1,650,000,000,000 total "human experience years"
  - *See my dataset linked at the bottom of this post.*
- 7,450,000,000 human years spent having sex
  - *Humans spend 0.45% of our lives having sex. 0.45% * [total human experience years] = 7E9 years*
- 52,000,000,000 years spent drinking coffee
  - *500 billion cups of coffee drunk this year x 15 minutes to drink each cup x 100 years* = 5E10 years*

        - **Coffee consumption has likely been much higher recently than historically, but it does have a long history. I'm estimating about a hundred years of current consumption for total global consumption ever.*
- 1,000,000,000 years spent in labor
    - *110,000,000,000 billion humans ever x ½ women x 12 pregnancies* x 15 hours apiece = 1.1E9 years*
        - **Infant mortality, yo. H/t Ellie and Shaw for this estimate.*
- 417,000,000 years spent worshipping the Greek gods
    - *1000 years* x 10,000,000 people** x 365 days a year x 1 hour a day*** = 4E8 years*
        - **Some googling suggested that people worshipped the Greek/Roman Gods in some capacity from roughly 500 BC to 500 AD.*
        - ***There were about 10 million people in Ancient Greece. This probably tapered a lot to the beginning and end of that period, but on the other hand worship must have been more widespread than just Greece, and there have been pagans and Hellenists worshiping since then.*
        - ****Worshiping generally took about an hour a day on average, figuring in priests and festivals? Sure.*
- 30,000,000 years spent watching Netflix
    - *14,000,000 hours/day* x 365 days x 5 years** = 2.92E7 years*
        - ** Netflix users watched an average of 14 million hours of content a day in 2017.*
        - ***Netflix the company has been around for 10 years, but has gotten bigger recently.*
- 50,000 years spent drinking coffee in Waffle House
    - *1.3 million cups* x 20 minutes = 4.9E4 years*
        - **Waffle House made this calculation easy. Bless you, Waffle House.*

So humanity in aggregate has spent about ten times as long worshiping the Greek gods as we've spent watching Netflix.

We've spent another ten times as long having sex as we've spent worshipping the Greek gods.

And we've spent ten times as long drinking coffee as we've spent having sex.

I'm not sure what this implies. Here are a few things I gathered from this:

1) I used to be annoyed at my high school world history classes for spending so much time on medieval history and after, when there was, you know, all of history before that too. Obviously there are other reasons for this - Eurocentrism, the fact that more recent events have clearer ramifications today - but to some degree this is in fact accurately reflecting how much history there is.

On the other hand, I spent a bunch of time in school learning about the Greek Gods, a tiny chunk of time learning about labor, and virtually no time learning about coffee. This is another disappointing trend in the way history is approached and taught, focusing on a series of major events rather than the day-to-day life of people.

2) The Funnel gets more stark the closer you move to the present day. Look at science. FLI reports that 90% of PhDs that have ever lived are alive right now. That

means most of all scientific thought is happening *in parallel* rather than *sequentially*.

3) You can't use the Funnel to reason about everything. For instance, you can't use it to reason about extended evolutionary processes. Evolution is necessarily cumulative. It works on the unit of generations, not individuals. (You can make some inferences about evolution - for instance, the likelihood of any particular mutation occurring increases when there are more individuals to mutate - but evolution still has the same number of generations to work with, no matter how large each generation is.)

4) This made me think about the phrase “living memory”. The world’s oldest living person is Kane Tanaka, who was born in 1903. 28% of the entirety of human experience has happened since her birth. As mentioned above, 15% has been directly experienced by living people. We have writing and communication and memory, so we have a flawed channel by which to inherit information, and experiences in a sense. But humans as a species can only directly remember as far back as 1903.

Here's my dataset. The population data comes from the Population Review Bureau and their report on how many humans ever lived, and from Our World In Data. Let me know if you get anything from this.

Fun fact: The average living human is 30.4 years old.

Wait But Why's explanation of the real revolution of artificial intelligence is relevant and worth reading. See also Luke Muehlhauser's conclusions on the Industrial Revolution: Part One and Part Two.

Coordination Problems in Evolution: Eigen's Paradox

Introduction

Lately I've written couple of posts that discuss coordination problems. Not the idealized, game-theoretical stuff but rather the real, messy coordination problems encountered by real people in the real world. Here, I will explore very different territory. I will look at coordination problems between between molecules, chromosomes, cells and individuals as they occurred and as they were solved in the course of biological evolution.

This article is based on the book "The Major Transitions in Evolution" by John Maynard Smith and Eörs Szathmáry.

Before proceeding I would like to say few words about why I chose that particular book, although it was published in 1995 and thus misses a lot of recent research.

First, it was written by widely recognized experts in the field. That may not have been that important if I was writing about a different topic, but evolutionary biology is notoriously tricky, subtle and prone to misunderstanding. Sometimes it generates crackpot ideas, which, nonetheless, sometimes turn out to be true. A layman, or even a popular science writer, is likely to get lost.

John Maynard Smith is one of the big names of evolutionary biology of 20th century. He owe to him the introduction of game theory into evolutionary biology. He's the author of the central idea in the field, so called evolutionarily stable strategy, which is, to put it shortly, an application of the concept of Nash equilibria to biological, evolving systems.

Eörs Szathmáry is less known, but he did a lot of work on the topic of origin of life.

Second, the book is concerned with the big changes in the evolutionary history. It doesn't spend much time on evolution-as-usual, on how a specific bone or organ evolved. Rather, it discusses the events which significantly changed the nature of evolution itself: How did the life began? How we've got the first self-replicating molecules? How did the cell originated? How did the multicellular organisms?

One would expect a book on such a grand subject to be a least a bit hand-wavy. Surprisingly though, it's not. Instead, the authors dive deep into the details of each individual topic, they discuss chemical details of the reactions in questions, their yield and speed, how would they survive in the competition of other reactions going on nearby and so on. They discuss the game-theoretic considerations of forming an eukaryotic cell or an insect society. They describe the minutiae of intragenomic conflict and the interplay between the development and evolution.

Third, in the introduction Smith and Szathmáry note that many (but not all) of the transitions they are going to discuss are, actually, solutions to coordination problems. They don't use that exact term, but it's pretty clear what they mean:

> One feature is common to many of the transitions: entities that were capable of independent replication before the transition can replicate only as part of a larger

> whole after it. … Given this common feature of the major transitions, there is a common question we can ask of them. Why did not natural selection, acting on entities at the lower level (replicating molecules, free-living prokaryotes, asexual protists, single cells, individual organisms), disrupt integration on the higher level (chromosomes, eukaryotic cells, sexual species, multicellular organisms, societies)?

In fact, thinking about coordination problems was what made them write the book in the first place:

> One of the stimuli for attempting the work was our realization that a model one of us had developed to analyze the origin of compartments containing populations of molecules was formally and mathematically similar to a model that the other had developed to analyze the evolution of cooperative behaviour in higher animals.

Fourth, the book strikes a good balance between targeting general public and targeting the experts only. It requires you to know your high school molecular and evolutionary biology, but not much more than that. You should be vaguely familiar with the concept of citric cycle, but nobody expects you to know what 1,3-biphosphoglycerate is. And once you know the basics, the book is surprisingly accessible and not hard to understand. (By the way, I see there's a pop version of the book published by the authors themselves. I haven't read it myself but it may be worth checking out.)

To sum it up, the book may be old, but it discusses exactly the topic I am interested in and it does so with great expertise and thoughtfulness. I don't think there's a newer book that does such a good job in this area.

And after all, my goal is not to summarize the cutting-edge biological research but to learn a lesson about the most general patterns of solving coordination problems. And those, I believe, haven't changed much in the past twenty years.

Eigen's paradox

How did the first self-replicating molecules originate?

We know that with RNA and some similar molecules this process happens automatically: If there are basic blocks available in the environment, they will, thanks to their chemical properties, automatically attach themselves into appropriate places of an existing single-strand RNA and form a double stranded RNA. For replication to proceed, the two strands then have to be separated. It have been proposed that this may have happened in the vicinity of hydrothermal vents, where the molecule would experience both cool temperatures, conductive to attachment of nucleotides to the RNA and sudden hot temperatures which would separate the two strands.

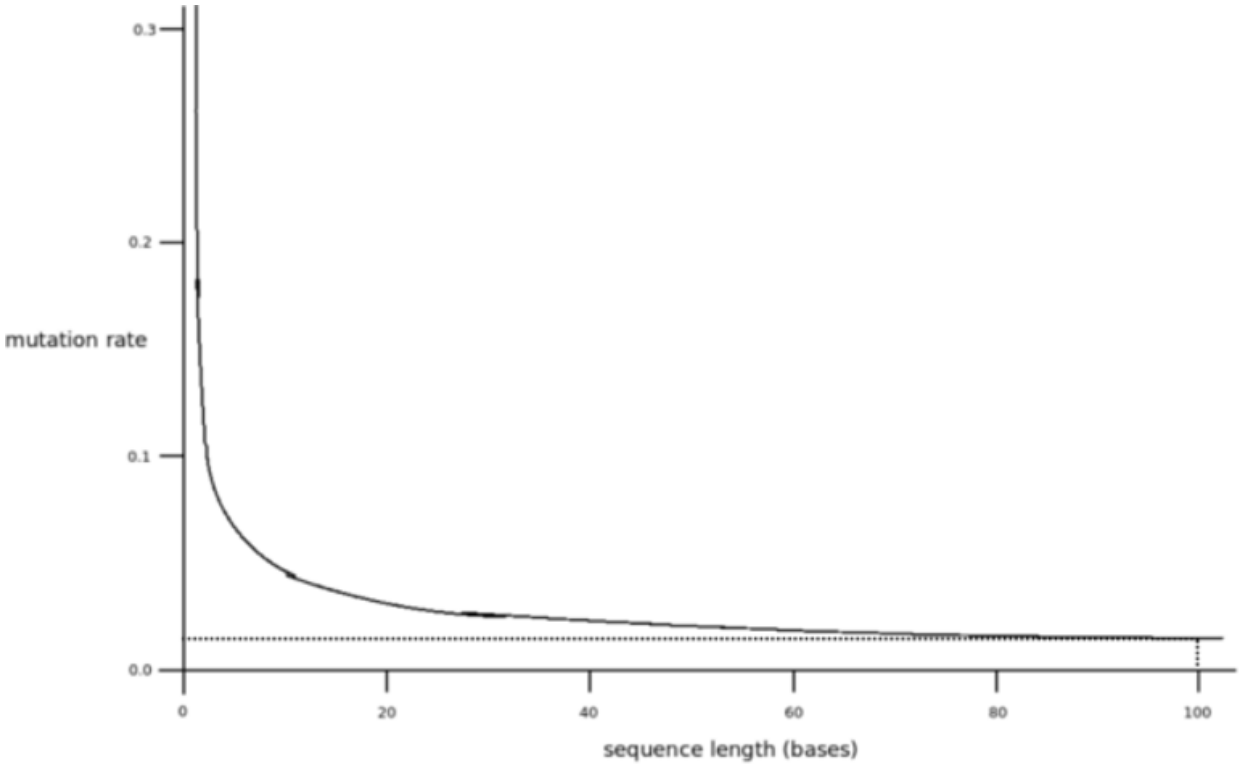
Now, putting aside the specifics of RNA replication, how likely it is that such a self-replicating molecule will survive in the chaotic world?

Manfred Eigen observes that it depends on two factors: On the speed of replication and on its fidelity. Speed allows molecule to be replicated faster than it decays. Fidelity ensures that the result of replication is, in fact, the molecule we care about and not something else.

As for the fidelity factor, it depends on the RNA size. If chance of correct replication of a single base is, for example, 1/2, then chance of correct replication of 2-base RNA is 1/4, 3-base RNA 1/8 and so on.

Measurements show that the threshold for the RNA size is somewhere around 100 bases. If the molecule is larger than that it wouldn't be able to sustain itself. It would devolve into a mix of its broken copies. (For the exact maths check the Wikipedia article.)

Following graph shows sustainable RNA size based on the mutation rate (1 - replication fidelity):



And here we encounter the paradox. If we wanted longer, self-sustainable RNA molecules we would need better fidelity. But better fidelity can only be achieved with the help of specialized enzymes. But the smallest genome able to code for such an enzyme, and for the necessary translating machinery, would require a number of bases greatly exceeding 100 nucleotides. It's a catch-22 situation.

Eigen speculates that the small molecules would have to somehow cooperate (Lo, a coordination problem!) to create a system capable of holding enough information to create more complex stuff. He proposes the "hypercycle" model.

Hypercycle is a set of RNA molecules that catalyze each other's replication. For example, if molecule A catalyses replication of molecule B, molecule B catalyses replication of molecule C and molecule C catalyses replication of molecule A, it's a hypercycle.

The point is that the catalysts can be specific. A should catalyze B and that's it. No need for complex machinery able to replicate any RNA sequence.

But before diving into the details: Could hypercycles even be established? Given the existence of small self-replicating RNA molecules, wouldn't the best replicator just crowd everyone else out and create a monoculture with no chance of forming a hypercycle?

Interestingly, no. And the reason is surprising. As RNA molecules replicate they have a tendency to pair with their own counterparts. (If fact, they bind so well that the problem is rather how the individual strands get separated after the replication.) And the bound double-stranded RNA cannot perform its catalytic functions. It inhibits its own creation.

In other words, the population of particular kind of molecule grows more slowly the more of its own kind is around. Its numbers do not double with each generation as one would naively expect. Yet, other kinds of molecules are not inhibited and can multiply at their own pace.

It can be shown that while in the world on exponentially replicating molecules the winner does take all, with sub-exponential growth, as described above, an equilibrium will form containing many different types of molecules.

Once we have the hypercycle in place it seems to work fine. Namely, notice how it is self-regulating in a way. If one link in the cycle is more efficient than other links it will soon run out of its catalyst and would have to wait for the rest of the cycle to catch up. Thus, a single component of the hypercycle cannot outcompete the rest of the cycle.

That being said, there's an obvious problem when we take mutations into account. If there is a mutation that makes one RNA molecule a less efficient catalyst for the next step of the cycle it could still reproduce at the same speed as the original molecule. That would mean lower concentration of the well-behaved molecule which would in turn suck the momentum out of the cycle. Many such free-riders and the concentration of cooperating RNAs would decrease to the level where the hypercycle would stop working at all.

So what can be done about the parasite molecules? An obvious solution would be to enclose the replicators in some kind of membrane. If the molecules in the compartment could replicate only together or not at all, the compartment containing parasites would simply "die" i.e. fail to replicate and be eventually outcompeted by the "healthy" compartments. (Emergence of membranes and the mechanism of the compartment fission is covered in the book but doesn't have much to do with the coordination problems, so I am going to hesitantly skip over it.) Alternatively, the evolution of the early life may have happened on a surface of a rock, thus limiting ways in which molecules can interact — this is so called "primordial pizza" model. Such anchoring of molecules to a flat surface may have had similar effect as enclosing them inside of a membrane.

Szathmáry and Demeter propose an alternative model called "stochastic corrector". The idea can be exemplified as follows.

Imagine a population of molecules consisting of "altruistic" molecules which catalyze replication and "parasites" which do not. Altruistic molecules reproduce less (it's hard to be both efficient catalyst and efficient replicator at the same time), the parasitic molecules reproduce more.

Let's assume that the molecules are either enclosed in compartments or tied to a surface in small patches. Each compartment or patch has to be small so that the law of big numbers doesn't kick in and make the proportion of the molecule types in all the compartments approximately the same.

The compartments with higher proportion of altruists are going to grow faster, the compartments with lower proportion of altruists are going to grow slower.

Then some external event, say a wave washing the molecules from the rock, mixes the molecules and creates a new arrangements of compartments or patches.

It can be shown that in such a setup there will arise a stable ratio of altruists and parasites. The parasites won't crowd out the altruists. To visualize the mechanism, imagine that every compartment contains only two RNA strands. The compartment containing two altruists will grow a lot. The compartment containing one altruist and one parasite will grow slower. Compartment with two parasites won't grow at all. After many iterations of mixing the molecules and repeating the process we'll arrive at a stable equilibrium of altruists and parasites.

Now think of compartments with three RNA strands. In that case it's more probable that any particular compartment will contain at least one parasite (7/8 as opposed to 3/4). The equilibrium will therefore contain more parasites than before. As we proceed to larger and larger compartments the relative advantage of parasites will grow until it reaches the point where the entire cycle will die off. It is therefore of essence that the compartment size remains small.

To sum it up, the stochastic corrector can work even without a hypercycle (there can be only one generalist catalyst molecule) but requires compartmentalization. Hypercycle, on the other hand, doesn't require compartments but is vulnerable to parasites. One can imagine a history where replication started with hypercycles and then, after membranes were formed, continued as a stochastic corrector.

The models above are, obviously, just a speculation. We don't have any remnants of those early stages of life and so guessing is the best we can do. Yet, some generic patterns, more examples of which we are going to encounter later, are beginning to emerge.

Chromosomes

At some point in the evolution the stand-alone genes stopped competing for themselves and started cooperating by getting linked into chromosomes.

Why would that be? Why link one's fate to other's rather than just keeping the status quo?

Apparently, linking into chromosomes comes with disadvantages for individual genes. Copying of the long linked RNA strands is slower. The conservative estimate of replication rate slow-down is 50%. That would make the linked gene severely disadvantaged in the competition with its stand-alone cousin. There has to be something that counterbalances that handicap.

The authors suggest that the main reason for this is what happens during the cell division.

If the daughter cell needs to contain all the genes to survive, it's crucial for any gene to end up in a cell that does contain at least one copy of each gene. Otherwise, no matter how successful they are individually, how many copies of themselves are present, they will end up in a dead end, trapped inside a non-functional cell.

This is especially true if the mechanism of the division is rather unsophisticated and probabilistic like, say, simple folding of the membrane and eventual random splitting of the content of the cell.

Furthermore, the more genes there are the more likely it is that one of them will be missing in the daughter cell. It may even happen that both daughter cells will miss a gene and die. This, I guess, places a hard upper limit on the number of stand-alone genes in the cell. With all genes linked in a single chromosome, on the other hand, it is much easier to succeed. The worst, though improbable, thing that can happen is that one of the daughter cells will end up with no genes at all.

To be continued

In the following parts of this article I would like to cover topics such as emergence of the eukaryotic cell, origin of mutli-celular life, of sex and of animal societies. In the end I am going to speculate about the very high-level, generic cooperation patterns and whether they have any semblance to coordination patterns that we encounter in human society.

Being a Robust Agent

Second version, updated for the 2018 Review. See change notes.

There's a concept which many LessWrong essays have pointed at it (indeed, I think the entire sequences are exploring). But I don't think there's a single post really spelling it out explicitly:

You might want to *become a more robust, coherent agent.*

By default, humans are a kludgy bundle of impulses. But we have the ability to reflect upon our decision making, and the implications thereof, and derive better overall policies.

Some people find this naturally motivating –it's aesthetically appealing to be a coherent agent. But if you don't find naturally appealing, the reason I think it's worth considering is *robustness* – being able to succeed at novel challenges in complex domains.

This is related to being instrumentally rational, but I don't think they're identical. If your goals are simple and well-understood, and you're interfacing in a social domain with clear rules, and/or you're operating in domains that the ancestral environment would have reasonably prepared you for… the most instrumentally rational thing might be to just follow your instincts or common folk-wisdom.

But instinct and common wisdom often aren't enough, such as when...

- You expect your environment to change, and default-strategies to stop working.
- You are attempting complicated plans for which there is no common wisdom, or where you will run into many edge-cases.
- You need to coordinate with other agents in ways that don't have existing, reliable coordination mechanisms.
- You expect instincts or common wisdom to be wrong in particular ways.
- You are trying to *outperform* common wisdom. (i.e. you're a maximizer instead of a satisficer, or are in competition with other people following common wisdom)

In those cases, you may need to develop strategies from the ground up. Your initial attempts may actually be worse than the common wisdom. But in the longterm, if you can acquire gears-level understanding of yourself, the world and other agents, you might eventually outperform the default strategies.

Elements of Robust Agency

I think of Robust Agency as having a few components. This is not exhaustive, but an illustrative overview:

- Deliberate Agency
- Gears-level-understanding of yourself
- Coherence and Consistency

- Game Theoretic Soundness

Deliberate Agency

First, you need to decide to be *any* kind of deliberate agent at all. Don't just go along with whatever kludge of behaviors that evolution and your social environment cobbled together. Instead, make conscious choices about your goals and decision procedures that you reflectively endorse,

Gears Level Understanding of Yourself

In order to reflectively endorse your goals and decisions, it helps to *understand* your goals and decisions, as well as intermediate parts of yourself. This requires many subskills, such as the ability to introspect, or to make changes to how your decision making works.

(Meanwhile, it also helps to understand how your decisions interface with the rest of the world, and the people you interact with. Gears level understanding is generally useful. Scientific and mathematical literacy helps you validate your understanding of the world)

Coherence and Consistency

If you want to lose weight and also eat a lot of ice cream, that's a valid set of human desires. But, well, it might just be impossible.

If you want to make long term plans that require commitment but also want the freedom to abandon those plans whenever, you may have a hard time. People you made plans with might get annoyed.

You can make deliberate choices about how to resolve inconsistencies in your preferences. Maybe you decide "actually, losing weight isn't that important to me", or maybe you decide that you want to keep eating all your favorite foods but also cut back on overall calorie consumption.

The "commitment vs freedom" example gets at a deeper issue – each of those opens up a set of broader strategies, some of which are mutually exclusive. How you resolve the tradeoff will shape what future strategies are available to you.

There are benefits to reliably being able to make trades with your future-self, and with other agents. This is easier if your preferences aren't contradictory, and easier if your preferences are either consistent over time, or at least *predictable* over time.

Game Theoretic Soundness

There are other agents out there. Some of them have goals orthogonal to yours. Some have common interests with you, and you may want to coordinate with them. Others may be actively harming you and you need to stop them.

They may vary in…

- What their goals are.
- What their beliefs and strategies are.
- How much they've thought about their goals.
- Where they draw their circles of concern.
- How hard (and how skillfully) they're trying to be game theoretically sound agents, rather than just following local incentives.

Being a robust agent means taking that into account. You must find strategies that work in a messy, mixed environment with confused allies, active adversaries, and sometimes people who are a little bit of both. (This includes creating credible incentives and punishments to deter adversaries from bothering, and motivating allies to become less confused).

Related to this is *legibility.* Your gears-level-model-of-yourself helps you improve your own decision making. But it *also* lets you clearly expose your policies to *other* people. This can help with trust and coordination. If you have a clear decision-making procedure that *makes sense*, other agents can validate it, and then you can tackle more interesting projects together.

Examples

Here's a smattering of things I've found helpful to think about through this lens:

- Be the sort of person that Omega can *clearly* tell is going to one-box – even a version of Omega who's only 90% accurate. Or, less exotically: Be the sort of person who your social network can clearly see is worth trusting, with sensitive information, or with power. Deserve Trust.
- Be the sort of agent who cooperates when it is appropriate, defects when it is appropriate, and can realize that cooperating-in-this-particular-instance might look superficially like defecting, but avoid falling into the trap.
- Think about the ramifications of people who think like you adopting the same strategy. *Not* as a cheap rhetorical trick to get you to cooperate on every conceivable thing. Actually think about how many people are similar to you. Actually think about the tradeoffs of worrying about a given thing. (Is recycling worth it? Is cleaning up after yourself at a group house? Is helping a person worth it? The answer *actually depends,* don't pretend otherwise).
- If there isn't enough incentive for others to cooperate with you, you may need to build a new coordination mechanism so that there *is* enough incentive. Complaining or getting angry about it *might* be a good enough incentive but often doesn't work and/or isn't quite incentivizing the thing you meant. (Be conscious of the opportunity costs of building *this* coordination mechanism instead of other ones. Be conscious of trying and failing to build a coordination mechanism. Mindshare is only so big)
- Be the sort of agent who, if some AI engineers were whiteboarding out the agent's decision making, they would see that the agent makes robustly good choices, such that those engineers would choose to implement that agent as software and run it.
- Be cognizant of order-of-magnitude. Prioritize (both for things you want for yourself, and for large scale projects shooting for high impact).
- Do all of this *realistically given your bounded cognition*. Don't stress about implementing a game theoretically perfect strategy, but *do* be cognizant how much computing power you actually have (and periodically reflect on whether

your cached strategies can be re-evaluated given new information or more time to think). If you're being simulated on a whiteboard right now, have at least a vague, credible notion of *how* you'd think better if given more resources.
- Do all of this realistically *given the bounded condition of *others*.* If you have a complex strategy that involves rewarding or punishing others in highly nuanced ways.... and they can't figure out what your strategy is, you may instead just be adding random noise instead of a clear coordination protocol.

Why is this important?

If you are a *maximizer*, trying to do something hard, it's hopefully a bit obvious why this is important. It's hard enough to do hard things *without* having incoherent exploitable policies and wasted motion chasing inconsistent goals.

If you're a satisficer, and you're basically living your life pretty chill and not stressing too much about it, it's less obvious that becoming a robust, coherent agent is useful. But I think you should at least consider it, because...

The world is unpredictable

The world is changing rapidly, due to cultural clashes as well as new technology. Common wisdom can’t handle the 20th century, let alone the 21st, let alone a singularity.

I feel comfortable making the claim: Your environment is *almost certainly* unpredictable enough that you will benefit from a coherent approach to solving novel problems. Understanding your goals and your strategy are vital.

There are two main reasons I can see to *not* prioritize the coherent agent strategy:

1. There may be higher near-term priorities.

You may want to build a safety net, to give yourself enough slack to freely experiment. It may make sense to first do all the obvious things to get a job, have enough money, and social support. (That is, indeed, what I did)

I'm not kidding when I say that building your decisionmaking from the ground up can leave you worse off in the short term. The valley of bad rationality be real, yo. See this post for some examples of things to watch out for.

Becoming a coherent agent is useful, but if you don't have a general safety net, I'd prioritize that first.

2. Self-reflection and self-modification is hard.

It requires a certain amount of mental horsepower, and some personality traits that not everyone has, including:

- Social resilience and openness-to-experience (necessary to try nonstandard strategies).
- Something like ‘stability’ or ‘common sense’ (I’ve seen some people try to rebuild their decision theory from scratch and end up hurting themselves).

- In general, the ability to think on purpose, and do things on purpose.

If you're the sort of person who ends up reading this post, I think you are probably the sort of person who would probably benefit (someday, from a position of safety/slack) from attempting to become more coherent, robust and agentic.

I've spent the past few years hanging around people who more agentic than me. It took a long while to really absorb their worldview. I hope this post gives others a clearer idea of what this path might look like, so they can consider it for themselves.

Game Theory in the Rationalsphere

That said, the reason I was motivated to write this wasn't to help individuals. It was to help with group coordination.

The EA, Rationality and X-Risk ecosystems include lots of people with ambitious, complex goals. They have many common interests and should probably be coordinating on a bunch of stuff. But they disagree on many facts, and strategies. They vary in how hard they've tried to become game-theoretically-sound agents.

My original motivation for writing this post was that I kept seeing (what seemed to me) to be strategic mistakes in coordination. It seemed to me that people were acting as if the social landscape was more uniform, and expecting people to be on the same "meta-page" of how to resolve coordination failure.

But then I realized that I'd been implicitly assuming something like "Hey, we're all *trying* to be robust agents, right? At least kinda? Even if we have different goals and beliefs and strategies?"

And that wasn't obviously true in the first place.

I think it's much easier to coordinate with people if you are able to model each other. If people have common knowledge of a shared meta-strategic-framework, it's easier to discuss strategy and negotiate. If multiple people are trying to make their decision-making robust in this way, that hopefully can constrain their expectations about when and how to trust each other.

And if you *aren't* sharing a meta-strategic-framework, that's important to know!

So the most important point of this post is to lay out the Robust Agent paradigm explicitly, with a clear term I could quickly refer to in future discussions, to check "is this something we're on the same page about, or not?" before continuing on to discuss more complicated ideas.

In praise of heuristics

We'll get there in the end, bear with me.

Introduction to ZD strategies in IPD

Feel free to skip if you're already familiar with ZD strategies.

In the iterated prisoner's dilemma (IPD) a zero determinant (ZD) strategy is one which forces your opponent's winnings to be a linear relation to your own winnings. These strategies can take either generous or extortionate forms.

Think of it as a modified version of tit-for-tat.

***

A Generous ZD strategy still always respond to C with C but will sometimes also respond to D with a C (it sometimes fails to punish). With "standard" PD utilities (T=0, R=-1, P=-2, S=-3) my opponent gains 1 utility by defecting. If I defect back in retaliation, I cost him 2 units of utility. If I defect back with probability 0.7, on average I cost him 1.4 units of utility. This still means that defecting is disadvantageous for my opponent (loss of 0.4 utility) but not quite as disadvantageous as it would be if I was playing pure tit-for-tat (loss of 1 utility).

This gets slightly more complex when you don't have constant gaps between T, R, P and S but the principle remains the same.

If he defects at all, my opponent will end up gaining more utility than me, but less than he would have got if he had co-operated throughout.

Advantages of GZD are:

1. Total utility isn't damaged as much by accidental defections as it is in pure tit-for-tat.

2. It won't get caught in endless C-D, D-C, C-D, D-C as tit-for-tat can.

***

On the other hand, Extortionate ZD always responds to D with D but also sometimes responds to C with D. Provided I don't respond to C with D too often, it is still advantageous to my opponent to play C (in terms of their total utility).

If my opponent co-operates at all I'll end up with more utility than him. If he gives in and plays C all the time (to maximise his own utility) I can achieve a better utility than I would with C-C. .

The main disadvantage of EZD in evolutionary games is that it defects against itself.

For both EZD and GZD you can vary your probabilities to be more or less generous/extortionate, provided you always ensure your opponent gets the most utility by co-operating.

Different opinions on fairness

An extortionate ZD strategy is similar to an opponent who has different perceptions of what is fair. Maybe your opponent had to pay to play the game but you got in free so he wants a higher percentage of the winnings. Maybe you think this is just his bad luck and think a 50:50 split is fair.

If you give in to what seems to you to be an extortionate strategy, your opponent is encouraged to make more extortionate demands in future, or modify his definition of what is fair. At some point, the level of extortion is so high that you barely get any advantage from co-operating.

This brings us to a proposal of Eliezer's.

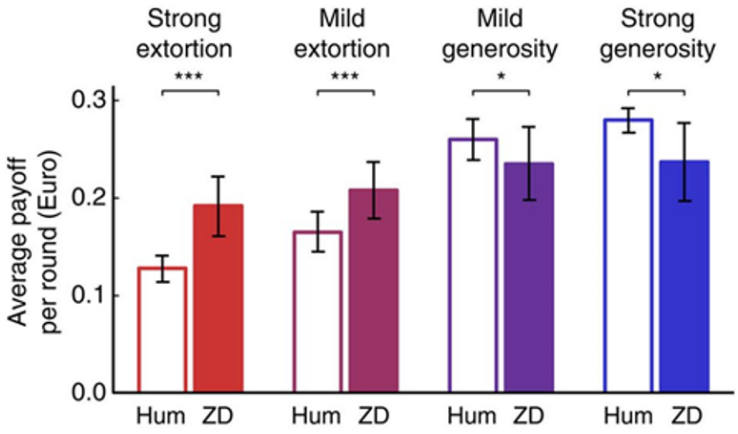
When choosing whether to give in to an extortioner you can capitulate to some extent, provided that you ensure your opponent gains less utility than he would if he agreed to your favoured position (ideally you should let your opponent know that this is what you're doing).

This removes any motivation to your opponent to extort and encourages him to give his true estimation of what is fair.

Two experiments in ZD strategies

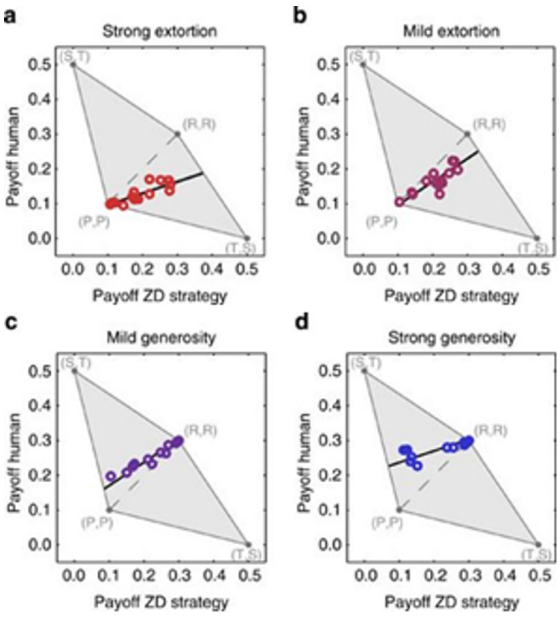
Hilbe et al. performed an experiment on humans playing against computerised ZD strategies. Four different strategies were tried – strong extortion through to strong generosity. Regrettably, pure tit-for-tat wasn't included as I would have liked to see a comparison with this.

The two generous strategies achieved higher average utility for the ZD programme than the 2 extortionate strategies. If the human players had acted purely in self-interest (they were paid according to the points gained) the extortionate strategists would have won. So what happened?



Firstly, a bit of detail about the experimental setup. The participants were not told that they were playing against a computer programme – the impression given was that they were playing against one of the other experimental subjects (although this wasn't explicitly stated).

Looking at the results from each individual player it is clear that none of the human participants allowed the extortionate ZD strategists to beat the score that is achievable from co-operating (R=0.3).



It seems that the subjects automatically used a strategy similar to the one suggest by Eliezer (or this represented something of a limit to co-operating) when dealing with a player who seemed to be extortionate.

In another experiment some players did allow the extortionate ZD strategy to achieve higher utility than it would have got by co-operating but over the two experiments there is a strong tendency not to let the unfair strategy get away with it.

***

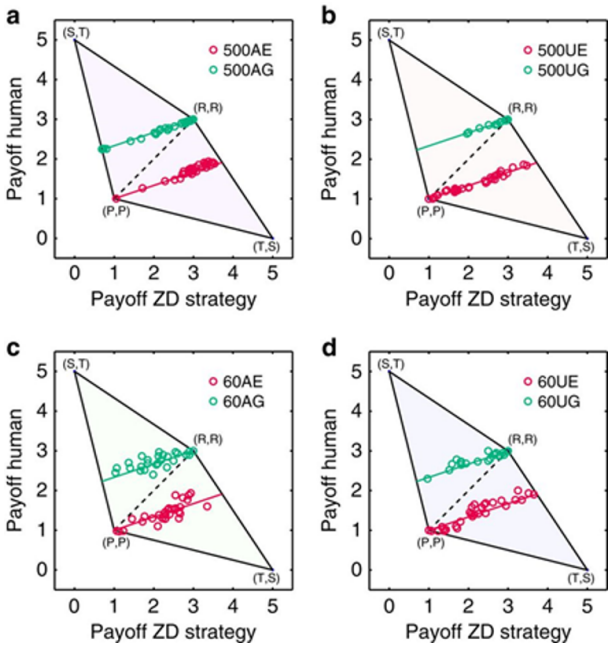
The second experiment tested the effects of:

1. More rounds of IPD (500 vs 60)

2. Being told that your opponent is a computer (Aware (A), Unaware (U))

3. Extortionate (E) / Generous (G) ZD strategies

Interestingly, human players in this second experiment were, over a long game, much more willing to let their EZD opponent “get away with it” when they were told that their opponent was a computer (see the grouping of red dots in figure a below). For the final 60 rounds of a 500 round IPD the extortionate ZD strategist was achieving 3.127 average utility – significantly (p=0.021) more than the R=3 gained from both players co-operating.



***

So why are we more willing to let a computer “get away with it”?

Maybe we view a computer as being insusceptible to change so are therefore more likely to give in.

Alternatively, if you think you are playing against a human, even after 500 rounds you will probably be annoyed enough with him for not co-operating properly that you won’t be co-operating the whole time. You’re less likely to get annoyed at a computer for beating you at a game. As soon as you realise you can’t beat the computer you can just try to do the best you can for yourself. This doesn't dent your pride as much as it would against a human opponent.

(There was one person who just defected pretty much throughout the whole 500AE experiment despite knowing he was playing against a computer, maybe he just decided to tit-for-tat or maybe it's just a lizardman constant. Without him the 500AE ZD strategy would have had even more impressive results.)

***

To me this looks like humans having a heuristic to deal with extortionate opponents/people who have a different opinion on what a fair split is. People seem to apply this heuristic naturally through emotions such as pride, annoyance and anger.

The heuristic works out roughly similarly to Eliezer’s suggestion of regulating your opponent’s winnings to less than he would achieve if he played fair/by your rules.

Being told that your opponent is a computer effectively turns off this heuristic (if you have long enough to get a rough idea of the computer’s strategy). This motivates your opponent to become more extortionate, something which the original heuristic was protecting you against.

In praise of heuristics

All of that is a very long introduction/example of my main point.

Heuristics are good.

Heuristics are very good.

You don’t even know how many times your heuristics have saved you.

You possibly have no idea what they are saving you from.

Knowing about biases can hurt people. Getting rid of heuristics without understanding them properly is potentially even more dangerous.

***

A recent discussion made me aware of this post by Scott where he tried to come up with a way of dealing with people who claim you have caused them offense.

One of the motivations for the post was Everybody Draw Mohammed Day. EDMD seems like it is a natural outworking of the heuristic described above.

People see the terrorists increasing their utility unfairly by attacking people who draw pictures of Mohammed. To ensure they don’t get an advantage by defecting, people want to decrease their utility back to below where they started – hence Everybody Draw Mohammed Day.

The terrorists were also following a similar heuristic – the original cartoon decreased their utility, they are trying to decrease the utility of those who created it to demotivate further defection.

The heuristic isn’t there to improve the world – it is just there so that the person performing it doesn’t encourage increased defection against themselves and increased demands from others.

Scott’s post was an attempt to turn off the heuristic and replace it with a principled position:

> The offender, for eir part, should stop offending as soon as ey realizes that the amount of pain eir actions cause is greater than the amount of annoyance it would take to avoid the offending action, even if ey can't understand why it would cause any pain at all. If ey wishes, ey may choose to apologize even though no apology was demanded

In this case, his proposal was criticised by others and Scott ended up rejecting his own proposal.

Had Scott applied his policy universally he would likely have ended up losing out if those he dealt with had modified to become more demanding of him.

It’s likely that our heuristic doesn’t lead us to an optimal result, it just prevents some bad results which Scott’s proposal may have led to.

Possibly a principled application of Eliezer’s proposal would help optimise the result better than the heuristic. In the experiments there was no standard amount that people chose to penalise the EZD strategist - the results were fairly spread out over the region between full defection and Eliezer's defined maximum co-operation. Sometimes the heuristic doesn't stop there and co-operates more than Eliezer would suggest.

***

This all sounds a bit harsh on Scott. Actually, putting an idea out there, engaging with criticism and admitting when you were wrong is exactly the right thing to do.

I’ll give an example where I didn’t do this and it did, in fact, end up biting me in the butt.

***

A while back, I was thinking about status. Status is, within a fixed group, a zero-sum game. People in the workplace are constantly attempting to improve their position on the ladder at the expense of others. This doesn’t just apply to promotions, it applies to pretty much everything. Alex wants to feel like he’s important and will get massively offended if he feels that Bob is trying to take status which should be Alex’s. This probably accounts for ~95% of disagreements in my workplace.

Zero-sum games are, usually, for suckers. If you can get out of the game and into a positive sum game, you probably should. This is doubly true if you’re competing for a thing you’re not really interested in.

Status very much matches my definition of a zero-sum game which I don’t want to play. The problem is, status also allows access to things which I do want – it is a very useful instrumental value. It is a game which everyone else plays so it’s hard to unilaterally leave.

Instead, I made the decision not to play status games unless I really have a need of the status (e.g. I will attempt to achieve status in the eyes of the person who will decide on a potential promotion but not others). Essentially I was trying to turn off the heuristic of “always attempt to gain status with everyone” and replace it with a trimmed down version “attempt to gain status only with those people who make a decision about your pay/promotions etc.”

Now if you have any experience in how status games work, you may realise that this was a naïve approach. If it isn’t obvious to you, have a think about what might go wrong.

*

*

*

If you don’t fight for your status with your colleagues, it’s like blood in the water. If they can push themselves up at your expense they will, not always maliciously, it’s just “the thing to do” in a workplace. If there are no consequences then it will happen again and again. In the end, this will mean that the people who you care about

impressing will see the status that others treat you as having and start to modify their own opinion of your status.

It took me a while to realise just how harmful this was. When I did, I had to do a lot of firefighting to re-establish a sense of normality.

All is fine again now but the experience did teach me that my heuristics are there for a reason and that I shouldn’t get rid of them entirely without properly understanding the consequences.

***

I haven’t decided exactly how I should deal with tackling heuristics in future but I have a few initial thoughts.

1. Don’t be overconfident that you have really understood why the heuristic is there

2. When comparing potential pros and cons, remember the cons are likely to be worse than you think

3. Discuss ideas with others

4. Where possible, make small changes first and monitor progress

Debate Rules In Benjamin Franklin's Junto

(Note: The Junto was a secret society formed by Benjamin Franklin for the purpose of intellectual discourse and business networking. The following is the debate rules they used to maintain an atmosphere of reason.)

1. Our debates were to be under the direction of a president, and to be conducted in the sincere spirit of inquiry after truth, without fondness for dispute, or desire of victory.
2. To prevent warmth, all expressions of positiveness in opinions, or direct contradiction, were after some time made contraband, and prohibited under small pecuniary penalties.
3. I even forbid myself, agreeably to the old laws of our Junto, the use of every word or expression in the language that imported a fix'd opinion, such as certainly, undoubtedly, etc., and I adopted, instead of them, I conceive, I apprehend, or I imagine a thing to be so or so; or it so appears to me at present.
4. When another asserted something that I thought an error, I deny'd myself the pleasure of contradicting him abruptly, and of showing immediately some absurdity in his proposition; and in answering I began by observing that in certain cases or circumstances his opinion would be right, but in the present case there appear'd or seem'd to me some difference, etc. I soon found the advantage of this change in my manner; the conversations I engag'd in went on more pleasantly. The modest way in which I propos'd my opinions procur'd them a readier reception and less contradiction; I had less mortification when I was found to be in the wrong, and I more easily prevail'd with others to give up their mistakes and join with me when I happened to be in the right.

Source: Excerpts from the Autobiography Of Benjamin Franklin

On Doing the Improbable

(Cross-posted from Facebook.)

I've noticed that, by my standards and on an Eliezeromorphic metric, most people seem to require catastrophically high levels of faith in what they're doing in order to stick to it. By this I mean that they would not have stuck to writing the Sequences or HPMOR or working on AGI alignment past the first few months of real difficulty, without assigning odds in the vicinity of 10x what I started out assigning that the project would work. And this is not a kind of estimate you can get via good epistemology.

I mean, you can legit estimate 100x higher odds of success than the Modest and the Outside Viewers think you can possibly assign to "writing the most popular HP fanfiction on the planet out of a million contenders on your first try at published long-form fiction or Harry Potter, using a theme of Harry being a rationalist despite there being no evidence of demand for this" blah blah et Modest cetera. Because in fact Modesty flat-out doesn't work as metacognition. You might as well be reading sheep entrails in whatever way supports your sense of social licensing to accomplish things.

But you can't get numbers in the range of what I estimate to be something like 70% as the required threshold before people will carry on through bad times. "It might not work" is enough to force them to make a great effort to continue past that 30% failure probability. It's not good decision theory but it seems to be how people actually work on group projects where they are not personally madly driven to accomplish the thing.

I don't want to have to artificially cheerlead people every time I want to cooperate in a serious, real, extended shot at accomplishing something. Has anyone ever solved this organizational problem by other means than (a) bad epistemology (b) amazing primate charisma?

EDIT: Guy Srinivasan reminds us that paying people a lot of money to work on an interesting problem is also standardly known to help in producing real perseverance.

List of previous prediction market projects

This is a linkpost for https://jacoblagerros.wordpress.com/list-of-prediction-markets/

Here I try to curate what is, to my knowledge, the most extensive list of previous prediction market projects.

This is useful because it provides a reference class to answer the question: “Why is our society inadequate at building functional futures markets for basically anything beyond equities, currencies and commodities, despite the massive benefits this would bring (see e.g. Arrow et al., 2008)?”

In this vein, I aim to primarily focus on real-money projects that might have scaled to something like an institutionalised exchange (and thus focus less on play-money projects, consulting companies offering a more nebolous “collective intelligence”, or forecasting platforms which are not actually *markets*).

I post this spreadsheet publicly despite its incompleteness, as doing so is likely the best way to complete it.

Feedback is very welcome.

If you -- yes you! -- are keen to help, the single most useful 15-30 min task would be to further research a cell in the column “Shut-down date and reason”, and comment with links/findings.

Genomic Prediction is now offering embryo selection

This is a linkpost for https://genomicprediction.com/epgt/

Book review: The Complacent Class

This is a linkpost for http://thinkingcomplete.blogspot.com/2018/08/book-review-complacent-class.html

THE COMPLACENT CLASS

The Self-Defeating Quest for the American Dream

TYLER COWEN

NEW YORK TIMES BESTSELLING AUTHOR OF *THE GREAT STAGNATION*

"*The best lack all conviction, while the worst*

Are full of passionate intensity."

--W. B. Yeats

The idea that things aren't going great these days is pretty widespread; there's a glut of books pointing out various problems. Cowen's achievement in this one is in weaving together disparate strands of evidence to identify the zeitgeist which summarises the overall trend - in a word, complacency. There are at least two ways in which people can be complacent. Either they're living pretty good lives, and want to solidify their positions as much as possible. Or they're unsatisfied with their lives, but unwilling to mobilise or take the risks which could improve their situations. (People in the middle of the economic spectrum showcase aspects of both). What's the opposite of complacency? Dynamism and risk-taking - traits which have always been associated with immigrants, and with America, the land of immigrants. Such traits aren't always expressed in positive ways, of course. Says Cowen: "Our current decade can be understood by comparing it to the 1960s and early 1970s. The Watts riots of 1965 put 4,000 people in jail and led to thirty-four killed and hundreds injured; during an eighteen-month period in 1971–1972, there were more than 2,500 domestic bombings reported, averaging out to more than five a day. I'm not advocating these tactics, of course. My point is that, today, there is an entirely different mentality, a far more complacent one, and one that finds it hard to grasp that change might proceed on such a basis. Yet in the 1960s and 1970s, not only did riots and bombings happen, but large numbers of influential intellectuals endorsed them, defended them, and maybe led them to some degree." In comparison to this, even people who are passionate about social change lack anywhere near the same sense of urgency today.

Cowen identifies many metrics which support the narrative of increasing complacency. Over the last few decades, interstate migration has gone down, job mobility has gone down, startup formation has gone down, and business churn and turnover have gone down. What's gone up? Market concentration and "matching": our ability to tailor our lives so that we're only exposed to things we're already comfortable with, whether that be music and movies similar to those we've seen before, or partners from the same class background as us. One particularly perverse result of better matching is increasing racial, economic and political segregation, reversing the progress made in the first half of the 20th century. A particularly notable cause of this is NIMBY movements, many of which have succeeded in ossifying their neighbourhoods by stifling new development. Segregation is also very pronounced in the incarceration industry.

In our personal lives, complacency involves prioritising comfort and security above all: physical security, with games like dodgeball and even tag being banned in schools; emotional security, via safe spaces and trigger warnings; and even corporate security, with companies hyper-focused on protecting their brands and other intangible assets (which have gone from less than 20% to over 80% of the value of the S&P 500 over the last 40 years). LGBT activists have moved from pushing the boundaries of societal norms to pursuing the most traditional of institutions, marriage. Nobody really has a bold vision for the future. (Relatedly, the portion of the federal budget allocated for discretionary spending has been falling sharply.)

But, Cowen says, this mindset can't keep limping along indefinitely, and will eventually face a crisis. In fact, we can think of it as a cyclic process: our current appetite for calm was whetted in the riotous and violent 70s and 80s - but as people

become more disillusioned, it will eventually give way to similar turmoil, the start of which we're already seeing.

Causes and complacencies

It's instructive to evaluate *The Complacent Class* with reference to Cowen's last two books. In *The Great Stagnation*, he argues that America's growth has slowed down because it ran out of "low-hanging fruit" like the technological advances of the early 20th century. Further, the problem is even worse than it seems, because growth in sectors like healthcare, finance and government spending contributes less to people's welfare than it used to. In *Average is Over,* he predicts the effects of the next low-hanging fruit: AI. He argues that those who can work well with technology - in broad strokes, the smart and the conscientious - will be able to replace dozens of less-skilled workers, and will be richly rewarded for it (which contributes to increasing credentialism). We should expect to see even more middle-class jobs crowded out, and even more wealth captured by a smaller proportion of people. Returns to being based in a good location are also increasing, driving the clustering of university graduates into relatively few cities. Meanwhile, Cowen predicts that the poor will be squeezed into places with much lower housing costs, even if they end up resembling "shantytowns". He notes that America's population is ageing, and that the elderly voting demographic usually gets what it wants, even at the cost of society overall.

Okay, so how does this relate to complacency? I think that's very unclear. If many of the trends cited in *The Complacent Class* can be explained by the ideas that we're in a "great stagnation" and that "average is over", without reference to individual attitudes, then they're not really evidence for complacency in the psychological sense. In fact, how do we know that there aren't other explanations for all of them? The ageing population comes to mind as a major possibility (although I'm not sure how many of Cowen's statistics already control for age). Perhaps it's useful to identify an overall pattern of "complacent" behaviour even if different changes have different causes, some psychological and some from technological and international shifts - but this sort of pattern has very little predictive power, since we don't know which other domains we can apply it to.

This lack of clarity around what complacency actually means makes it somewhat unfalsifiable. For Cowen, complacency is demonstrated both by the rich erecting barriers to the advancement of the poor, and by the poor not breaking through those barriers. But what if the rich did nothing while their social and financial dominance eroded away - wouldn't that also count as complacency? What if the poor are actually willing to take more risks now (like the financial risk of going to university) but it's just paying off less - would that really make them "complacent"? Cowen claims that more relaxed codes of dress and manners display "a culture of the static and the settled", but don't they also lower implicit class barriers and therefore promote dynamism? There's a case that doing graduate degrees is ambitious and valuable, but there's a similarly strong case that it's a complacent replacement for making things happen in the real world. The very same companies which match us with our preferred options also allow us to sample more variety - whether that's in songs, shopping, or sexual partners. And so on. More generally, I think we should be biased against claims of the form "it's a big problem that the next generation have the wrong attitude towards X", because they have occurred so commonly throughout history, usually sounded convincing, and were usually wrong.

Nevertheless, there's undeniably some truth to Cowen's core argument. Almost none of the physical technologies around us (buildings, cars, trains, rockets, household appliances) have seen significant progress over the last half-century; nor have systems like healthcare, law, politics or education. But more importantly, people aren't even surprised by this stasis: the radical expectations of the mid-20th century have given way to doubt that our lives will be any better than our parents', plus a generous helping of political disillusionment. It's true that IT has made massive leaps, but as Cowen notes, "a lot of the internet’s biggest benefits are distributed in proportion to our cognitive abilities to exploit them". People who don't highly value near-unlimited access to information or niche communities may even find that the downsides of the internet (addiction to games or porn, mental health problems exacerbated by social media, news media's race to the bottom) outweigh its upsides. So I do believe that westerners today are, psychologically, more complacent than they were a few decades ago, and that this shows through in attitudes towards risk and expectations for the future. It's also likely that increasingly complacent behaviour which isn't caused by a complacent mentality still leads to an overall culture of complacency, although disentangling cause and effect here is tricky. Either way, Cowen's ideas are as thought-provoking as usual and should be taken into account by anyone interested in understanding America.

An Undergraduate Reading Of: Semantic information, autonomous agency and non-equilibrium statistical physics

This is a recent paper by Artemy Kolchinsky and David H. Wolpert, from the Santa Fe Institute. It was published in The Royal Society Interface on Oct 19. They propose a formal theory of semantic information, which is to say how to formally describe meaning. I am going over it in the style proposed here and shown here, approximately.

I will go through the sections in-line at first, and circle back if appropriate. Mostly this is because when I pasted the table of contents it very conveniently kept the links to the direct sections of the paper, which is an awesome feature.

- Abstract: as is my custom, skipped.
- 1. Introduction
    - *Semantic* information is meaningful to a system, as distinct from *syntactical* information, which is correlational.
    - They cite the importance of the idea in these fields: biology, cognitive science, artificial intelligence, information theory, and philosophy.
    - Question one: can it be defined formally and generally?
    - Question two: can that definition be used in any physical system (rocks, hurricanes, cells, people)?
    - They claim the answer to both questions is yes. They define *semantic information:*

> the information that a physical system has about its environment that is causally necessary for the system to maintain its own existence over time

    - Most of the time we study syntactic information, using Shannon's theory.
    - Shannon explicitly avoided addressing what meaning a message over a telecommunication line might have.
    - One approach to address this is to assume an idealized system that optimizes some function, e.g. utility.
    - Under this approach, semantic information helps the system achieve its goal.
    - The problem with it is the goal is defined exogenously. Therefore meaning is to the scientists who impute goals to the system, not the system itself.
    - We want meaning based on the intrinsic properties of the system.
    - In biology the goal of an organism is fitness maximization, which leads to the *teleosemantic* approach, which roughly says a trait has meaning if at some time in the past it correlated with states of the environment (and therefore had a bearing on fitness).
    - Example: frogs snap their tongues at black spots in their visual field. This is semantic information because eating flies is good for frogs, and correlated with flies in the past.
    - The problem with teleosemantics is it defines meaning in terms of the past history of the system; an ahistorical definition that relies only on the dynamics of the system in a given environment is the goal.

- Another approach is *autonomous agents*, which maintain their own existence in an environment. This has self-preservation as the goal, and does not rely on history.
- Autonomous agents get information about the environment, and then respond in 'appropriate' ways. Example:

For instance, a chemotactic bacterium senses the direction of chemical gradients in its particular environment and then moves in the direction of those gradients, thereby locating food and maintaining its own existence.

- Research suggests the information used for self-maintenance is meaningful, but this concept has remained informal. In particular, there is no formal way to quantify the semantic information an agent has, or to determine the meaning of a particular state.
- Their contribution:

We propose a formal, intrinsic definition of semantic information, applicable to any physical system coupled to an external environment

- A footnote here says that the method should generalize to any dynamical system, but they focus on physical ones in the paper. This is an interesting claim to me.
- There is ‘the system X’ and ‘the environment Y’; at some initial time $t = 0$, they are jointly distributed according to some initial distribution $p(x0, y0)$; they undergo coupled (possibly stochastic) dynamics until time τ, where τ is some timescale of interest.
- There is a *viability function*, which is the negative Shannon entropy of the distribution over the states of system X. This quantifies the 'degree of existence' at any given time. More information about the viability function in section 4.
- Shannon entropy is used because it provides an upper bound on the probability of states of interest; it also has a well-developed connection to thermodynamics, which links them to non-equilibrium statistical physics.
- Semantic information is a subset of syntactic information which causally contributes to the continued existence of the system. This maintains the value of the viability function.
- They draw from Pearl, and use a form of interventions in order to quantify:

To quantify the causal contribution, we define counter-factual **intervened distributions** in which some of the syntactic information between the system and its environment is scrambled.

- Figure 1 has some graphical examples.
- They give three verbal examples for the scrambling procedure: switching rocks between fields, switching hurricanes between oceans, and switching birds between environments. This section is a little suspect to me; the hurricane and the rock were described as "low viability value of information" when scrambling consisted of putting them in very similar environments, but then the bird was "high viability value of information" when scrambling put them in *random* environments. Further, the rock and bird were on year timelines, and the hurricane only an hour. This might just be sloppy explanation. In the main, I would expect the lifespan of a system to be inversely correlated with viability value of information overall, so I would have thought hurricane>bird>rock.

    - They use 'coarse-graining' methods from information theory to formalize transforming the actual distribution into intervened distributions.
    - The intervention which has the same (or greater?) viability as the actual distribution, but has the least syntactical information, is called the *viability-optimal intervention*.
    - They interpret all of the syntactic information of the optimal intervention to be semantic information, because any further scrambling changes the viability.
    - *Semantic efficiency* is the ratio of semantic to syntactic information. It quantifies how tuned the system is to only gather information relevant to its existence.
    - *Semantic content* of a system state x is the conditional distribution, under the optimal intervention, given state x. This can tell us the correlations relevant to maintaining the system.
    - They claim to be able to do point-wise semantic information as well.
    - The framework is not tied to the Shannon notion of syntactic information; from different measures of syntactic information they can derive appropriate measures of semantic information, e.g. thermodynamics through statistical physics.
    - Measures of semantic information are defined relative to the choice of:

(1) the particular division of the physical world into ‘the system’ and ‘the environment’;

(2) the timescale τ; and

(3) the initial probability distribution over the system and environment.

    - They suggest implications for an intrinsic definition of autonomous agency.

- 2. Non-equilibrium statistical physics: Body - skipped for now.
- 3. Preliminaries and physical set-up: Body - skipped for now.
- 4. The viability function: Body - mostly skipped, but I did go in to find the actual function:
    - define the viability function as the negative of the Shannon entropy of the marginal distribution of system x at time τ,
    - $V(p_{X_r}) := -S(p_{X_r}) = \sum_{x_r} p(x_\tau) \log p(x_\tau)$
- 5. Semantic information via interventions: Body - skipped for now.
- 6. Automatic identification of initial distributions, timescales and decompositions of interest: Body - skipped for now.
- 7. Conclusion and discussion
    - Semantic information is syntactic information that is causally necessary for the system to continue.
    - It can be stored (mutual information between system and environment) and observed (transfer entropy exchanged between system and environment).
    - Semantic information can misrepresent the world. This shows up as a negative viability value.
    - Semantic information is asymmetrical between system and environment.
    - No need to decompose the system into different degrees of freedom (sensors/effectors, body/brain, membrane/interior).
    - Side-steps the question of internal models or representations entirely.

    - The framework does not assume organisms, but it may be useful for offering quantitative and formal definitions of life.
    - They suggest that high semantic information may be a necessary, though not sufficient, condition for being alive.

Note: I have left the links below for completeness, and to make it easy to interrogate the funding/associations of the authors. The appendices have some examples they develop.

End: I am putting this up before delving into the body sections in any detail, not least for length and readability. If there is interest, I can summarize those in the comments.

Fasting Mimicking Diet Looks Pretty Good

Epistemic status: pretty much factual.

CW: diets, calories

One of the odd things about working on longevity is that now people ask me for lifestyle advice.

Or they ask me what *I, personally* do to live longer.

Mostly my response has been a lame “um, nothing?”

There are, as of now, *no* interventions shown to make humans live longer or slow or reverse the human aging process. And, of the interventions reported to make animals live longer, many are doubtful, and many are too risky or unpleasant to make the cost-benefit tradeoff look good for healthy people.

Also, as a personal matter, I’m just not very interested in my own “lifestyle optimization” for the most part. My motivation is about helping people, not especially staving off death for myself; I think I’m more mentally prepared for death than most people my age. Certainly I’ve thought about it more concretely. (BTW, if you too like to know all the gory and technical details about how people die, this blog by an ICU nurse is gold.)

And “lifestyle optimization” turns out to be heavily about diet and exercise, and…I confess, diet culture *really creeps me out*. Not at all my thing.

That said, there *is* a lifestyle intervention that seems pretty evidence-based and also pretty low on risk and inconvenience: the *Fasting Mimicking Diet*, developed by Valter Longo of USC.

It’s actually been tested in a clinical trial on 100 healthy participants, where it improved a bunch of biomarkers related to aging and disease (reduced IGF and blood pressure, though no change in glucose, triglycerides, cholesterol, or CRP.)

The really good results are in mice, where it rescues both Type I and Type II diabetes as well as a mouse model of MS, reduces tumors by 45% and dermatitis by 50%, increases mesenchymal stem cells by 45x, improves motor and cognitive performance, and results in an 11% lifespan extension.

So, what is the FMD?

It’s a 5-day low-calorie, low-carb, low-protein diet, followed by a period of eating however you would by default.

Caloric restriction (reducing calorie intake about 1/3 from baseline or ad-lib) is probably the most replicated lifespan- and healthspan-extending intervention in animals. It’s about 30-40% life extension in mice and rats. In monkeys, it extends lifespan little if at all, but delays age-related disease and hair loss. However, the side effects are nontrivial — humans on CR experience weakness, lethargy, depression,

muscle wasting, and neurological deficits. (Undereating also stunts growth in children and adolescents, and underweight in women causes infertility, miscarriage, and preterm birth.)

Mice seem to get most of the benefits of CR, including an equally extended lifespan, from an isocaloric but low-protein or low-methionine diet. Low-protein diets are safe for humans and might not be as damaging to quality of life, but they do definitely inhibit physical fitness/performance.

Alternate-day fasting in mice has a bunch of benefits, including lifespan extension of 10-30% depending on mouse strain, as well as reduction in cancer incidence, and lower levels of neural damage in mouse models of Alzheimer's, Huntington's, Parkinson's, and acute brain injury. In a randomized controlled trial in humans, alternate-day fasting caused weight loss but no improvement in metabolic/cardiovascular parameters.

The FMD seems like the *least* amount of dietary restriction that is still known to cause life extension. 5 days/month of low calorie intake isn't that big a commitment.

Valter Longo sells patented packaged foods for the FMD, but they're pricey ($300 for five days).

What I find more aesthetic, and cheaper, is an adapted version, which I'm trying now:

For the first five weekdays of every month, eat nothing but (non-potato) vegetables, cooked in fat if desired. The rest of the time, eat whatever you want.

It's low-calorie and low-protein while containing vitamins, but it skips the calorie-counting and allows you to actually cook tasty food.

Since I'm breastfeeding, which is about a 500-calorie daily expenditure, it's a little harder on me than it would be by default, so I'm adding the modification of *if you feel weak or lightheaded, eat a fat source until you stop feeling that way.* I expect this is probably a good conservative measure for people in general.

This ought to be generally safe for healthy adults under 65. The clinical trial reported no adverse effects more serious than fatigue.

It's definitely *not* a good idea for children, diabetics, pregnant people, or people with disordered eating.

If you basically believe the science that periods of little or no food promote good metabolic processes (autophagy, reduced inflammation, increased neurogenesis & stem cell production) but you don't want the nasty side effects of prolonged caloric restriction, *some* kind of intermittent or periodic fasting seems like a sensible thing to try.

I don't think there's any direct evidence that the FMD is better than intermittent fasting for health, but it seems easier to do, and maybe a bit better in terms of results from randomized human trials.

If you (like me) really don't like the aesthetics of dieting — "special" pre-packaged foods, appearance insecurity, calorie counting, having to make excuses to the people around you for eating "weirdly" — a homebrew FMD is pretty ideal because you are spending very little time "on a diet", and you are eating normal things (vegetables).

Also, it’s not necessarily a weight-loss diet, and you can conceptualize it as primarily about health, not looks.

I *don’t* expect it to have nontrivial lifespan effects on humans, but it might be good for healthspan or disease risk, and that seems worthwhile to me.

Alignment Newsletter #28

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

Highlights

Motivating the Rules of the Game for Adversarial Example Research *(Justin Gilmer, George E. Dahl et al)* (summarized by Dan H): In this position paper, the authors argue that many of the threat models which motivate adversarial examples are unrealistic. They enumerate various previously proposed threat models, and then they show their limitations or detachment from reality. For example, it is common to assume that an adversary must create an imperceptible perturbation to an example, but often attackers can input whatever they please. In fact, in some settings an attacker can provide an input from the clean test set that is misclassified. Also, they argue that adversarial robustness defenses which degrade clean test set error are likely to make systems less secure since benign or nonadversarial inputs are vastly more common. They recommend that future papers motivated by adversarial examples take care to define the threat model realistically. In addition, they encourage researchers to establish “content-preserving” adversarial attacks (as opposed to “imperceptible” l_p attacks) and improve robustness to unseen input transformations.

Dan H's opinion: This is my favorite paper of the year as it handily counteracts much of the media coverage and research lab PR purporting ``doom'' from adversarial examples. While there are some scenarios in which imperceptible perturbations may be a motivation---consider user-generated privacy-creating perturbations to Facebook photos which stupefy face detection algorithms---much of the current adversarial robustness research optimizing small l_p ball robustness can be thought of as tackling a simplified subproblem before moving to a more realistic setting. Because of this paper, new tasks such as Unrestricted Adversarial Examples (AN #24) take an appropriate step toward increasing realism without appearing to make the problem too hard.

Technical AI alignment

Agent foundations

A Rationality Condition for CDT Is That It Equal EDT (Part 2) *(Abram Demski)*

Learning human intent

Learning under Misspecified Objective Spaces *(Andreea Bobu et al)*: What can you do if the true objective that you are trying to infer is outside of your hypothesis space? The key insight of this paper is that in this scenario, the human feedback that you get will likely not make sense for *any* reward function in your hypothesis space, which allows you to notice when this is happening. This is operationalized using a Bayesian model in which a latent binary variable represents whether or not the true objective is in the hypothesis space. If it is, then the rationality constant β will be large (i.e. the human appears to be rational), whereas if it is not, then β will be small (i.e. the human

appears to be noisy). The authors evaluate with real humans correcting the trajectory of a robotic arm.

Adversarial Imitation via Variational Inverse Reinforcement Learning *(Ahmed H. Qureshi et al)*: A short history of deep IRL algorithms: GAIL introduced the idea of training a policy that fools a discriminator that tries to distinguish a policy from expert demonstrations, GAN-GCL showed how to recover a reward function from the discriminator, and AIRL (AN #17) trains on (s, a, s') tuples instead of trajectories to reduce variance, and learns a reward shaping term separately so that it transfers better to new environments. This paper proposed that the reward shaping term be the *empowerment* of a state. The empowerment of a state is the maximum mutual information between a sequence of actions from a state, and the achieved next state. Intuitively, this would lead to choosing to go to states from which you can reach the most possible future states. Their evaluation shows that they do about as well as AIRL in learning to imitate an expert, but perform much better in transfer tasks (where the learned reward function must generalize to a new environment).

Rohin's opinion: I'm confused by this paper, because they only compute the empowerment for a *single action*. I would expect that in most states, different actions lead to different next states, which suggests that the empowerment will be the same for all states. Why then does it have any effect? And even if the empowerment was computed over longer action sequences, what is the reason that this leads to learning generalizable rewards? My normal model is that IRL algorithms don't learn generalizable rewards because they mostly use the reward to "memorize" the correct actions to take in any given state, rather than learning the underlying true reward. I don't see why empowerment would prevent this from happening. Yet, their experiments show quite large improvements, and don't seem particularly suited to empowerment.

Task-Embedded Control Networks for Few-Shot Imitation Learning *(Stephen James et al)*

Adversarial examples

Motivating the Rules of the Game for Adversarial Example Research *(Justin Gilmer, George E. Dahl et al)*: Summarized in the highlights!

Verification

Verification for Machine Learning, Autonomy, and Neural Networks Survey *(Weiming Xiang et al)*

Robustness

Iterative Learning with Open-set Noisy Labels *(Yisen Wang et al)* (summarized by Dan H): Much previous research on corrupted learning signals deals with label corruption, but this CVPR 2018 paper considers learning with corrupted or irrelevant inputs. For example, they train a CIFAR-10 classifier on CIFAR-10 data mixed with out-of-class CIFAR-100 data; such a scenario can occur with flawed data curation or data scraping. They use a traditional anomaly detection technique based on the local outlier factor to weight training examples; the more out-of-distribution an example is, the less weight

the example has in the training loss. This approach apparently helps the classifier cope with irrelevant inputs and recover accuracy.

Making AI Safe in an Unpredictable World: An Interview with Thomas G. Dietterich *(Thomas G. Dietterich and Jolene Creighton)*

Read more: Open Category Detection with PAC Guarantees is the corresponding paper.

Miscellaneous (Alignment)

Standard ML Oracles vs Counterfactual ones *(Stuart Armstrong)*: *(Note: This summary has more of my interpretation than usual.)* Consider the setting where an AI system is predicting some variable y = f(x), but we will use the AI's output to make decisions that could affect the true value of y. Let's call the AI's prediction z, and have y = g(x, z), where g captures how humans use z to affect the value of y. The traditional ML approach would be to find the function f that minimizes the distance between y_i and f(x_i) on past examples, but this does not typically account for y depending on z. We would expect that it would converge to outputting a fixed point of g (so that y = z = g(x, z)), since that would minimize its loss. This would generally perform well; while manipulative predictions z are possible, they are unlikely. The main issue is that since the system does not get to observe z (since that is what it is predicting), it cannot model the true causal formulation, and has to resort to complex hypotheses that approximate it. This can lead to overfitting that can't be simply solved by regularization or simplicity priors. Instead, we could use a counterfactual oracle, which reifies the prediction z and then outputs the z that minimizes the distance between z and y, which allows it to model the causal connection y = g(x, z).

Rohin's opinion: This is an interesting theoretical analysis, and I'm surprised that the traditional ML approach seems to do so well in a context it wasn't designed for. I'm not sure about the part where it would converge to a fixed point of the function g, I've written a rambly comment on the post trying to explain more.

Misbehaving AIs can't always be easily stopped! *(El Mahdi El Mhamdi)*

AI strategy and policy

The Future of Surveillance *(Ben Garfinkel)*: While we often think of there being a privacy-security tradeoff and an accountability-security tradeoff with surveillance, advances in AI and cryptography can make advances on the Pareto frontier. For example, automated systems could surveil many people but only report a few suspicious cases to humans, or they could be used to redact sensitive information (eg. by blurring faces), both of which improve privacy and security significantly compared to the status quo. Similarly, automated ML systems can be applied consistently to every person, can enable collection of good statistics (eg. false positive rates), and are more interpretable than a human making a judgment call, all of which improve accountability.

China’s Grand AI Ambitions with Jeff Ding *(Jeff Ding and Jordan Schneider)*

On the (In)Applicability of Corporate Rights Cases to Digital Minds *(Cullen O’Keefe)*

Other progress in AI

Exploration

Episodic Curiosity through Reachability *(Nikolay Savinov, Anton Raichuk, Raphael Marinier, Damien Vincent et al)* (summarized by Richard): This paper addresses the "couch potato" problem for intrinsic curiousity - the fact that, if you reward an agent for observing novel or surprising states, it prefers to sit in front of a TV and keep changing channels rather than actually exploring. It proposes instead rewarding states which are difficult to reach from already-explored states (stored in episodic memory). Their agent has a separate network to estimate reachability, which is trained based on the agent's experiences (where observations few steps apart are negative examples and those many steps apart are positive examples). This method significantly outperforms the previous state of the art curiousity method on VizDoom and DMLab environments.

Richard's opinion: This paper is a useful advance which does help address the couch potato problem, but it seems like it might still fail on similar problems. For example, suppose an agent were given a piece of paper on which it could doodle. Then states with lots of ink are far away from states with little ink, and so it might be rewarded for doodling forever (assuming a perfect model of reachability). My guess is that a model-based metric for novelty will be necessary to counter such problems - but it's also plausible that we end up using combinations of techniques like this one.

Reinforcement learning

Open Sourcing Active Question Reformulation with Reinforcement Learning *(Michelle Chen Huebscher et al)*: Given a question-answering (QA) system, we can get better performance by reformulating a question into a format that is better processed by that system. (A real-world example is google-fu, especially several years ago when using the right search terms was more important.) This blog post and accompanying paper consider doing this using reinforcement learning -- try a question reformulation, see if gives a good answer, and if so increase the probability of generating that reformulation. For this to work at all, the neural net generating reformulations has to be pretrained to output sensible questions (otherwise it is an *extremely* sparse reward problem). They do this by training an English-English machine translation system. The generated reformulations are quite interesting -- 99.8% start with "what is name", and many of them repeat words. Presumably the repetition of words is meant to tell the underlying QA system that the word is particularly important.

Rohin's opinion: I like how this demonstrates the faults of our current QA systems -- for example, instead of understanding the semantic content of a question, they instead focus on terms that are repeated multiple times. In fact, this might be a great way to tell whether our systems are "actually understanding" the question (as opposed to, say, learning a heuristic of searching for sentences with similar words and taking the last noun phrase of that sentence and returning it as the answer). For a good QA system, one would hope that the optimal question reformulation is just to ask the same question again. However, this won't work exactly as stated, since the RL system could learn the answers itself, which could allow it to "reformulate" the question such that the answer is obvious, for example reformulating "In what year did India gain independence?" to "What is 1946 + 1?" Unless the QA system is perfectly

optimal, there will be some questions where the RL system could memorize the answer this way to improve performance.

Learning Acrobatics by Watching YouTube *(Xue Bin (Jason) Peng et al)*: To imitate human behavior in videos, it is sufficient to estimate the human pose for each frame, to smooth the poses across frames to eliminate any jittery artifacts or mistakes made by the pose estimator, and then to train the robot to match the motion exactly. This results in really good performance that looks significantly better than corresponding deep RL approaches, but of course it relies on having labeled poses to train the pose estimator in addition to the simulator.

Rohin's opinion: It's quite remarkable how some supervision (poses in this case) can lead to such large improvements in the task. Of course, the promise of deep RL is to accomplish tasks with very little supervision (just a reward function), so this isn't a huge breakthrough, but it's still better than I expected. Intuitively, this works so well because the "reward" during the imitation phase is extremely dense -- the reference motion provides feedback after each action, so you don't have to solve the credit assignment problem.

Reinforcement Learning for Improving Agent Design *(David Ha)*: This paper explores what happens when you allow an RL agent to modify aspects of the environment; in this case, the agent's body. This allows you to learn asymmetric body designs that are better suited for the task at hand. There's another fun example of specification gaming -- the agent makes its legs so long that it simply falls forward to reach the goal.

Meta learning

CAML: Fast Context Adaptation via Meta-Learning *(Luisa M Zintgraf et al)*

Unsupervised learning

Unsupervised Learning via Meta-Learning *(Kyle Hsu et al)* (summarized by Richard): This paper trains a meta-learner on tasks which were generated using unsupervised learning. This is done by first learning an (unsupervised) embedding for a dataset, then clustering in that embedding space using k-means. Clustering is done many times with random scaling on each dimension; each meta-learning task is then based on one set of clusters. The resulting meta-learner is then evaluated on the actual task for that dataset, performing better than approaches based just on embeddings, and sometimes getting fairly close to the supervised-learning equivalent.

Richard's opinion: This is a cool technique; I like the combination of two approaches (meta-learning and unsupervised learning) aimed at making deep learning applicable to many more real-world datasets. I can imagine promising follow-ups - e.g. randomly scaling embedding dimensions to get different clusters seems a bit hacky to me, so I wonder if there's a better approach (maybe learning many different embeddings?). It's interesting to note that their test-time performance is sometimes better than their training performance, presumably because some of the unsupervised training clusterings are "nonsensical", so there is room to improve here.

Applications

Learning Scheduling Algorithms for Data Processing Clusters *(Hongzi Mao et al)*

Miscellaneous (AI)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *(Jacob Devlin et al)*

PPO-CMA: Proximal Policy Optimization with Covariance Matrix Adaptation *(Perttu Hämäläinen et al)*

News

Internships and fellowships for 2019: There are a lot of AI internships and fellowships to apply for now, including the CHAI summer internship (focused on safety in particular), the OpenAI Fellows, Interns and Scholars programs, the Google AI Residency Program (highlights), the Facebook AI Research Residency Program, the Microsoft AI Residency Program, and the Uber AI Residency.

The AAAI's Workshop on Artificial Intelligence Safety

What will the long-term future of employment look like?

This is a linkpost for http://thinkingcomplete.blogspot.com/2018/02/what-will-long-term-future-of.html

Current debates over the automation of human jobs by AI tend to focus on how fast it will proceed: some jobs will be safe for the next decade, some for the next half-century, some for the foreseeable future. In this essay I want to take a different approach, by asking which (if any) jobs will continue to be done by humans even after the development of arbitrarily advanced technology.

[Here I've omitted a long section on why AI is capable of becoming arbitrarily competent; it was relevant to the original competition into which this essay was entered, but it's nothing new to anyone on this site. See my linked blog for the full post and references].

The fact that AI will eventually utterly outclass humans in performance on any given information-processing task does not mean that humans will necessarily be shut out from those industries: by Ricardian comparative advantage, it could still be profitable to employ humans even when they have no absolute advantage over AI. [6] However, since copying software is trivially easy, and the price of hardware continues to plummet, such roles will be very limited, especially if there is any sector of the economy in which humans have a comparative advantage over AI.

I claim that there is such a sector, consisting of what I shall call 'social jobs'. I define these as jobs where most or all of the value produced comes not from their outputs, but from the fact that they are being performed by other humans. The principle is most clearly illustrated by personal relationships: we want friends and partners who not only say and do the right things, but also really feel love and respect for us. In the economic sphere, there are particularly important social components of jobs in sales, management and leadership, and entertainment. To understand them, it's instructive to consider jobs whose value is almost entirely tied to the fact that humans are involved. Take chess, for instance. If we view the value of professional chess players as due to their production of high-quality games, then the advent of superhuman chess engines has rendered them entirely redundant. Yet professional chess not only still exists, but also features ever-growing prize pools. [7] So we are better to model chess players as entertainers, whose humanity is central to the existence of their jobs.

There are plenty more examples to consider. Fans of sports matches or rock concerts could often see and hear the performances better if they watched a broadcast from home; however, as with most entertainment, it's the atmosphere and sense of connection with the people around them that they value more. A robot could say and do all the same things as a lobbyist or politician, but still wouldn't thereby build the personal relationships that are crucial for both jobs. An AI counselor might be much more knowledgeable about psychology than a human one, and yet be much less able to convince the patient that someone actually cares about them. In all these cases, the outputs produced by an AI may be very similar to or better than a human's, but we would assign the latter much more value, because we care that there is a person on the other end of each interaction. In fact, humans will be able to do such jobs much

better by incorporating analysis from AIs into their judgements - but human participation will still be essential.

Growth in these roles, and the social economy overall, will be driven by massive increases in overall wealth. Piketty estimated that world GDP increased by a factor of 140 between 1700 and 2012. [8] Historically, such gains have led to significantly increased spending on entertainment, and the rapid expansion of intrinsically social sectors like fashion and tourism. [9] Such spending is the most likely to support irreplaceable human jobs in the long term, for the reasons I outlined above - and there are good reasons to think that the current "Fourth Industrial Revolution" will be of even greater scale than its predecessors, allowing these social industries to grow rapidly. [10]

It may seem that, even so, there will simply not be enough social jobs for everyone. However, it’s unlikely that the current dichotomy between employment and unemployment will last indefinitely: technology is very good at introducing flexibility to previously rigid systems. Amazon's Mechanical Turk, for instance, is a marketplace where people can be paid for doing various tasks whenever they want, without any commitment or fixed hours. Driving for Uber or Lyft is the same. There are also now many games where you can trade virtual goods gained during gameplay for real money. These particular opportunities will eventually be automated away, but there will be plenty of replacements - and if global wealth increases to the extent suggested above, then such part-time jobs will be sufficient to maintain a reasonable quality of life.

In fact, we are already seeing surges in flexible social jobs, driven by the ease of connecting with people online. "Social media influencers" such as Instagram stars can earn millions; so can motivational speakers and life coaches with popular Youtube videos. [11] The massive earning power of media figures and celebrities is not a new phenomenon, but it's important to note the decreasing barriers to entry. Anyone can set up an Instagram account and start amassing followers (even if right now only a few can monetise effectively). Meanwhile, websites like Patreon, Kickstarter and meetup.com allow grassroots entrepreneurs and content creators to reach larger audiences (and profits) while remaining responsive to their supporters.

The ability of technology to facilitate peer-to-peer interactions suggests the potential for social jobs to become far more bottom-up (as opposed to the top-down model epitomised by TV culture). For example, during the last century an incredibly broad array of subcultures - from board game enthusiasts to queer communities to the worldwide competitive debating circuit - have largely been run by unpaid volunteers. However, if more wealth will indeed be funnelled into entertainment and leisure as the AI revolution progresses, and tools for distributing this money continue to be developed, then the economies of these communities will become more complex, with more opportunities to translate passion for a subculture into a decent wage. This is exactly what happened in many sports over the last century as participation and viewership boomed. Now millions of people are employed in sports-related jobs - and many such roles will never be automated away while the sport lasts, because their social aspects are pivotal in tying their respective communities together, and because they confer prestige within each community.

The exact activities and achievements considered prestigious will, of course, be mediated by cultural factors. For example, domestic workers make up less than 1% of the workforce in developed countries but up to a dozen times that in other parts of the world. [12] A significant portion of this disparity can be attributed to cultural reticence

towards hiring "servants" in the West. However, in general, consuming goods or services is more prestigious the more scarce or expensive they are. For example, in very poor countries possession of factory-made "Western" products can be a proud distinction. Yet in countries where automated production of goods is commonplace, it is the opposite: markets such as Etsy are able to charge a premium for handmade goods. In a future where goods and services can be produced incredibly cheaply by automation, the ability to hire other people to work for you will remain a scarce resource, and therefore will in general become more of a signal of social status. Fast food outlets, which compete mostly on price, lead the restaurant industry in automation; whereas the fanciest restaurants and hotels, which compete mostly on reputation, will likely always keep their human waiters, chefs, valets, receptionists and concierges.

The predictions I have outlined above depend on people still caring about the distinction between humans and AI. I think this is likely. Friendships and relationships are fundamentally based on a mutual emotional bonds and shared concsious experiences between equals. Of course, consciousness is mysterious, and perhaps we will create AIs who (we are convinced) consciously feel emotions in the same way as humans. However, unless those AIs are specifically engineered to be very similar to humans, there will be certain facets of human relationships that they can't replicate. Could you feel like an equal in a friendship with a being who thinks ten thousand times faster than you, living a lifetime during each of your days? Or wholly trust a partner intelligent enough to manipulate you in any way they desired? I don't want to say that such shifts will never occur, but my best guess is that even in the very long term, most people will want their personal relationships to continue involving other humans. And as long as these non-transactional experiences are highly prized, then it will be economically valuable for human workers to evoke emotions that AIs do not. Even if distinguishing AIs from humans becomes difficult in some cases (e.g. via convincing facial and vocal simulations), it also seems likely that legal measures will be taken to prevent AIs from masquerading as human.

Finally, note that arguments above should not be taken to apply specifically to the coming few decades. Economically speaking, it is much more difficult to model transitions between equilibria than those equilibria themselves. While it seems very likely that most current jobs will eventually be replaced by the sorts of social jobs I described above, the exact path taken will depend on how fast technology advances, and how governments respond; the last few years have demonstrated that both of these factors are very tricky to predict. However, even when the most complex technical skills have become redundant, as long as the fundamentals impulses of human nature remain, there will still be an economy driven by humans and our social interactions.

The Kelly Criterion

Epistemic Status: Reference Post / Introduction

The Kelly Criterion is a formula to determine how big one should wager on a given proposition when given the opportunity.

It is elegant, important and highly useful. When considering sizing wagers or investments, if you don't understand Kelly, *you don't know how to think about the problem.*

In almost every situation, reasonable attempts to use it will be somewhat wrong, but superior to ignoring the criterion.

What Is The Kelly Criterion?

The Kelly Criterion is defined as (from Wikipedia):

> For simple bets with two outcomes, one involving losing the entire amount bet, and the other involving winning the bet amount multiplied by the payoff odds, the Kelly bet is:
>
> 
>
> where:
>
> - f^* is the fraction of the current bankroll to wager, i.e. how much to bet;
> - b is the net odds received on the wager ("b to 1"); that is, you could win $b (on top of getting back your $1 wagered) for a $1 bet
> - p is the probability of winning;
> - q is the probability of losing, which is $1 - p$.
>
> As an example, if a gamble has a 60% chance of winning ($p = 0.60$, $q = 0.40$), and the gambler receives 1-to-1 odds on a winning bet ($b = 1$), then the gambler should bet 20% of the bankroll at each opportunity ($f^* = 0.20$), in order to maximize the long-run growth rate of the bankroll.

(A bankroll is the amount of money available for a gambling operation or series of wagers, and represents what you are trying to grow and preserve in such examples.)

For quick calculation, you can use this rule: bet such that you are trying to win a percentage of your bankroll *equal to your percent edge.* In the above case, you win 60% of the time and lose 40% on a 1:1 bet, so you on average make 20%, so try to win 20% of your bankroll by betting 20% of your bankroll.

Also worth remembering is if you bet *twice* the Kelly amount, on average the geometric size of your bankroll *will not grow at all,* and anything larger than that will on average cause it to shrink.

If you are trying to grow a bankroll that cannot be replenished, Kelly wagers are an *upper bound* on what you can ever reasonably wager, and 25%-50% of that amount is the sane range. You should be *highly suspicious* if you are considering wagering anything above half that amount.

(Almost) never go full Kelly.

Kelly betting, or betting full Kelly, is correct if *all* of the following are true:

1. You care only about the long-term geometric growth of your bankroll.
2. Losing your entire bankroll would indeed be infinitely bad.
3. You do not have to worry about fixed costs.
4. When opportunities to wager arise, you never have a size minimum or maximum.
5. There will be an unlimited number of future opportunities to bet with an edge.
6. You have no way to meaningfully interact with your bankroll other than wagers.
7. You can handle the swings.
8. You have full knowledge of your edge.

At least seven of these eight things are *almost never* true.

In most situations:

1. Marginal utility is decreasing, but in practice falls off far less than geometrically.
2. Losing your entire bankroll would end the game, but that's life. You'd live.
3. Fixed costs, including time, make tiny bankrolls only worthwhile for the data and experience.
4. There is *always* a maximum, even if you'll probably never hit it. Long before that, costs go up and people start adjusting the odds based on your behavior. If you're a small fish, smaller ponds open up that are easier to win in.
5. There are only so many opportunities. Eventually we are all dead.
6. At some cost you can usually earn money and move money into the bankroll.
7. You can't handle the swings.
8. You don't know your edge.

There are two reasons to preserve one's bankroll. A bankroll provides opportunity to get data and experience. One can use the bankroll to make money.

Executing real trades is necessary to get worthwhile data and experience. Tiny quantities work. A small bankroll with this goal must be preserved and variance minimized. Kelly is far too aggressive.

If your goal is profit, $0.01 isn't much better than $0.00. You'll need to double your stake seven times to even have a dollar. That will take a long time with 'responsible' wagering. The best thing you can do is bet it all long before things get that bad. If you lose, you can walk away. Stop wasting time.

Often you should do both simultaneously. Take a small amount and grow it. Success justifies putting the new larger amount at risk, failure justifies moving on. One can say that this can't possibly be optimal, but it is simple, and psychologically beneficial, and a limit that is easy to justify to oneself and others. This is often more important.

The last reason, #8, is the most important reason to limit your size. If you often have less edge than you think, but still have some edge, reliably betting too much will often turn you from a winner into a loser. Whereas if you have *more* edge than you think, and you end up betting too little, that's all right. You're gonna be rich anyway.

For compactness, I'll stop here for now.

A Rationality Condition for CDT Is That It Equal EDT (Part 2)

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

The previous post sketched an application of Jessica's COEDT framework to get rid of one of the assumptions of my argument for CDT=EDT. Looking at the remaining assumptions of my argument, the hardest one to swallow was implementability: the idea that when the agent implements a mixed strategy, its randomization is successfully controlling for any factors other than those involved in the decision to use that particular mixed strategy. Stated in bayes-net terms, the action has no parents other than the decision.

I stated that the justification of the assumption had to do with the learnability of the causal connections. I then went on to discuss some issues in learning counterfactuals, but not in a way which directly addressed the implementability assumption.

The present post discusses aspects of learning which are more relevant to the implementability assumption. Actually, though, these considerations are only arguments for implementability in that they're arguments for CDT=EDT. None of the arguments here are watertight.

Whereas the previous post was largely a response to COEDT, this post is more exclusively talking about CDT=EDT.

How Should We Learn Counterfactuals?

When Are Counterfactuals Reasonable?

How do we evaluate a proposed way to take logical counterfactuals? One reason progress in this area has been so slow is that it has been difficult to even state *desirable properties* of logical counterfactuals in a mathematically concrete way, aside from strong intuitions about how specific examples should go.

However, there is *one* desirable property which is very clear: ***counterfacting on what's true should result in what's true***. You might get a strange alternative system of mathematics if you ask what things would be like if 2+2=3, but counterfacting on 2+2=4 just gets us mathematics as we know it.

When we consider situations where an agent fully understands what decision problem it's in (so we are assuming that its map corresponds perfectly to the territory), this means that counterfacting on the action which it really takes in that situation, the counterfactual should tell it the true consequences of that action. However, it is less clear how to apply the principle for learning agents.

In the purely Bayesian case, one might argue that the prior is the equivalent of the true circumstance or decision problem; the best an agent can do is to respond as if it were put in situations randomly according to its prior. However, this doesn't give any guarantees about performance in particular situations; in particular, it doesn't give guarantees about learning the right counterfactuals. This becomes a somewhat more pressing concern when we move beyond Bayesian cases to logical induction, since there isn't a prior which can be treated as the true decision problem in the same way.

One might argue, based on the scary door problem from the previous post, that we shouldn't worry about such things. However, I think there are reasonable things we can ask for without going all the way to opening the scary door.

I propose the following principle for learning agents: ***you should not be able to systematically correct your counterfactuals.*** The principle is vaguely stated, but my intention is that it be similar to the logical induction criterion in interpretation: there shouldn't be efficient corrections to counterfactuals, just like there shouldn't be efficient corrections to other beliefs.

If we run into sufficiently similar situations repeatedly, this is like the earlier truth-from-truth principle: the counterfactual predictions for the action actually taken should not keep differing from the consequences experienced. The principle is also not so strong that it implies opening the scary door.

An Example

Consider a very unfair game of matching pennies, in which the other player sees your move before playing. In the setup of Sam's logical induction tiling result, where the counterfactual function can be given arbitrarily, the most direct way to formalize this makes it so that the action takes always gets utility zero, but the *other* action always *would have yielded* utility one, if it had been taken.

The CDT agent doesn't know at the time of making the decision whether the payoff will be 0 for heads and 1 for tails, or 1 for heads and 0 for tails. It can learn, however, that which situation holds depends on which action it ends up selecting. This results in the CDT agent acting pseudorandomly, choosing whichever action it least expects itself to select. The randomization doesn't help, and the agent's expected utility for each action converges to 1/2, which is absurd since it always gets utility 0 in reality. This expectation of 1/2 can be systematically corrected to 0.

An EDT agent who understands the situation would expect utility 0 for each action, since it knows this to be the utility of an action if that action is taken. The CDT agent also knows this to be the conditional expected utility if it actually takes an action; it just doesn't care. Although the EDT agent and CDT agent do equally poorly in this scenario, we can add a third option to not play, giving utility 1/4. The EDT agent will keep going after the illusory 1/2 expected utility.

Sam's theorem assumes that we aren't in a situation like this. I'm not sure if it came off this way in Alex's write-up, but Sam tends to talk about cases like this as unfair environments: the way the payoff for actions depends on the action we actually take makes it impossible to do well from an objective standpoint. I would instead tend to say that this is a bad account of the counterfactuals in the situation. If the other player can see your move in matching pennies, you're just going to get 0 utility no matter how you move.

If we think of this as a hopelessly unfair situation, we accept the poor performance of an agent here. If we think of it as a problem with the counterfactuals, we want a CDT agent to be arranged internally such that it couldn't end up with counterfactuals like these. Asking that counterfactuals not be systematically correctible is a way to do this.

It looks like there is already an argument for EDT along similar lines in the academic literature: *The Gibbard-Harper Collapse Lemma for Counterfactual Decision Theory.*

Can We Dutch Book It?

We can get a CDT agent to bet against the correctible expectations if we give it a side-channel to bet on which doesn't interfere with the decision problem. However, while this may be peculiar, it doesn't really constitute a Dutch Book against the agent.

I suspect a kind of Dutch Book could be created by asking it to bet in the same act as its usual action (so that it is causally conditioning on the action at the time, and therefore choosing under the influence of a correctable expectation). This could be combined with a side-bet made after choosing. I'm not sure of the details here, though.

EDIT: Yes, we can dutch-book it.

A Condition for Reflective Stability of CDT Is That It Equal EDT

Sam's Result

Perhaps a stronger argument for requiring CDT counterfactuals to equal EDT conditionals is the way the assumption seems to turn up in expected-value tiling results. I mentioned earlier that Sam's logical induction tiling result uses a related assumption. Here's Alex Appel's explanation of the assumption:

> The key starting assumption effectively says that the agent will learn that the expected utility of selecting an action in the abstract is the same as the expected utility of selecting the specific action that it actually selects, even if a finite-sized chunk of the past action string is tampered with. Put another way, in the limit, if the agent assigns the best action an expected utility of 0.7, the agent will have its expected utility be 0.7, even if the agent went off-policy for a constant number of steps beforehand.

The part about "even if a finite-sized chunk of the past action string is tampered with" is not so relevant to the present discussion. The important part is that the expected utility of a specific action is the same as that expected without knowing which action will be selected.

One way that this assumption can be satisfied is if the agent learns to accurately predict which action it will select. The expectation of this action equals the expectation in general because the expectation in general already knows this action will be taken.

Another way this assumption can be satisfied is if the counterfactual expectation of each action which the agent might take equals the evidential expectation of that action. In other words, the CDT expectations equal the EDT expectations. This implies the desired condition because the counterfactual expectations of each action which the agent believes it might take end up being the same. (If there were a significant difference between the expectations, the agent could predict the result of its argmax, and therefore would know which action it will take.)

The assumption could also hold in a more exotic way, where the counterfactual expectations and the evidential expectations are not equal individually, but the differences balance each other out. I don't have a strong argument against this possibility, but it does seem a bit odd.

So, I don't have an argument here that CDT=EDT is a necessary condition, but it is a sufficient one. As I touched on earlier, there's a question of whether we should regard this as a property of fair decision problems, or as a design constraint for the agent. The stronger condition of knowing your own action isn't a feasible design constraint; some circumstances make your action unpredictable (such as "I'll give you $100 - [the degree of expectation you had of your action before taking that action]"). We can, however, make the counterfactual expectations equal the evidential ones.

Diff's Tiling Result

I don't want to lean heavily on Sam's tiling result to make the case for a connection between CDT=EDT and tiling, though, because the conclusion of Sam's theorem is that an agent won't envy another agent for its *actions*, but *not* that it won't self-modify into that other agent. It might envy another agent for *counterfactual* actions which are relevant to getting utility. Indeed, counterfactual mugging problems don't violate any of the assumptions of Sam's theorem, and they'll make the agent want to self-modify to be updateless. Sam's framework only lets us examine sequences of actions, and whether the agent would prefer to take a sequence of actions other than its own. It doesn't let us examine whether the agent's own sequence of actions involves taking a screwdriver to its internal machinery.

Diff's tiling result is very preliminary, but it does suggest that the relationship between CDT=EDT and tiling stays relevant when we deal with self-modification. (Diff = Alex Appel) His first assumption states that concrete expected utility of an action takes equals abstract expected utility of taking some action, along very similar lines to the critical assumption in Sam's theorem.

We will have to see how the story plays out with tiling expected utility maximizers.

The Major Caveat

The arguments in this post are all restricted to *moves which the agent continues to take arbitrarily many times as it learns about the situation it is in.* The condition for learning counterfactuals, that they not be systematically correctible, doesn't mean anything for actions which you never take. You can't Dutch Book inconsistent-seeming beliefs which are all conditional on something which never happens. And weird beliefs about actions which you never take don't seem likely to be a big stumbling block for tiling results; the formalizations I used to motivate the connection had conditions having to do with actions actually taken.

This is a major caveat to my line of reasoning, because even if counterfactual expectations and conditional expectations are equal for actions which continue to be taken arbitrarily often, CDT and EDT may end up taking entirely different actions in the limit. For example, in XOR Blackmail, CDT refuses to respond to the letter, and EDT responds. Both expect disaster not to occur after their respective actions, and both are right in their utility forecasts.

We can use Jessica's COEDT to define the conditional beliefs for actions we don't take, but where is the argument that counterfactual beliefs must equal these?

We could argue that CDTs, too, should be limits of CDTs restricted to mixed strategies. This is very much the intuition between trembling-hand Nash equilibria. We then argue that CDTs restricted to mixed strategies should take the same actions as EDTs, by the arguments above, since there are no actions which are never taken. This argument might have some force if a reason for requiring CDT to be the limit of CDTs restricted to mixed strategies were given.

My intuition is that XOR blackmail really shouldn't be counted as a counterexample here. It strikes me as a case correctly resolved by UDT, not CDT. Like counterfactual mugging, counterfactual actions of the agent factor in to the utility received by the agent; or, putting it a different way, a copy of the agent is run with spoofed inputs (in contrast with Newcomb's problem, which runs an exact copy, no spoofing). This means that an agent reasoning about the problem should either reason updatelessly about how it should respond to that sort of situation, or doubt its senses (which can be equivalent).

In other words, my intuition is that there is a class of decision problems for which rational CDT agents converge to EDT in terms of actions taken, not only in terms of the counterfactual expectations of those actions. This class would nicely rule out problems requiring updateless reasoning, and include XOR Blackmail among them.

Two Kinds of Technology Change

Many of the examples in this post are drawn from Volume I of Fernand Braudel's "Civilization and Capitalism: 15th-18th Century", which I strongly recommend for anyone interested in a quantitative approach to history.

You may have heard that Gutenberg revolutionized the intellectual world with his invention of moving type in the mid-1400's.

Here's the thing, though: Gutenberg was not the first to try movable type. The Chinese were using basic printing presses in the ninth century; Pi Cheng introduced movable characters between 1040 and 1050. So why didn't it catch on then? And even setting that aside, surely *some* tired monk must have thought of the idea sooner.

Turns out, prior to the 14th century, books were primarily printed on parchment—created from sheep skins. A single 150-page book required the skins of 12 sheep to make the parchment. That much parchment wasn't cheap—the parchment on which a book was written cost far more than the actual writing. With that much cost sunk in the materials, it's no wonder that book-buyers wanted beautiful, handwritten script—it added relatively little to the cost.

It was paper which changed all that. European paper production didn't get properly underway until the 1300's. Once it did, book prices plummeted, writing became the primary expense of book production, and printing presses with movable type followed a century later.

The printing press offers a clear example of a technology change whose arrival was limited, not by the genius of the inventor, but by economic viability. The limiting factor wasn't insight, it was prices.

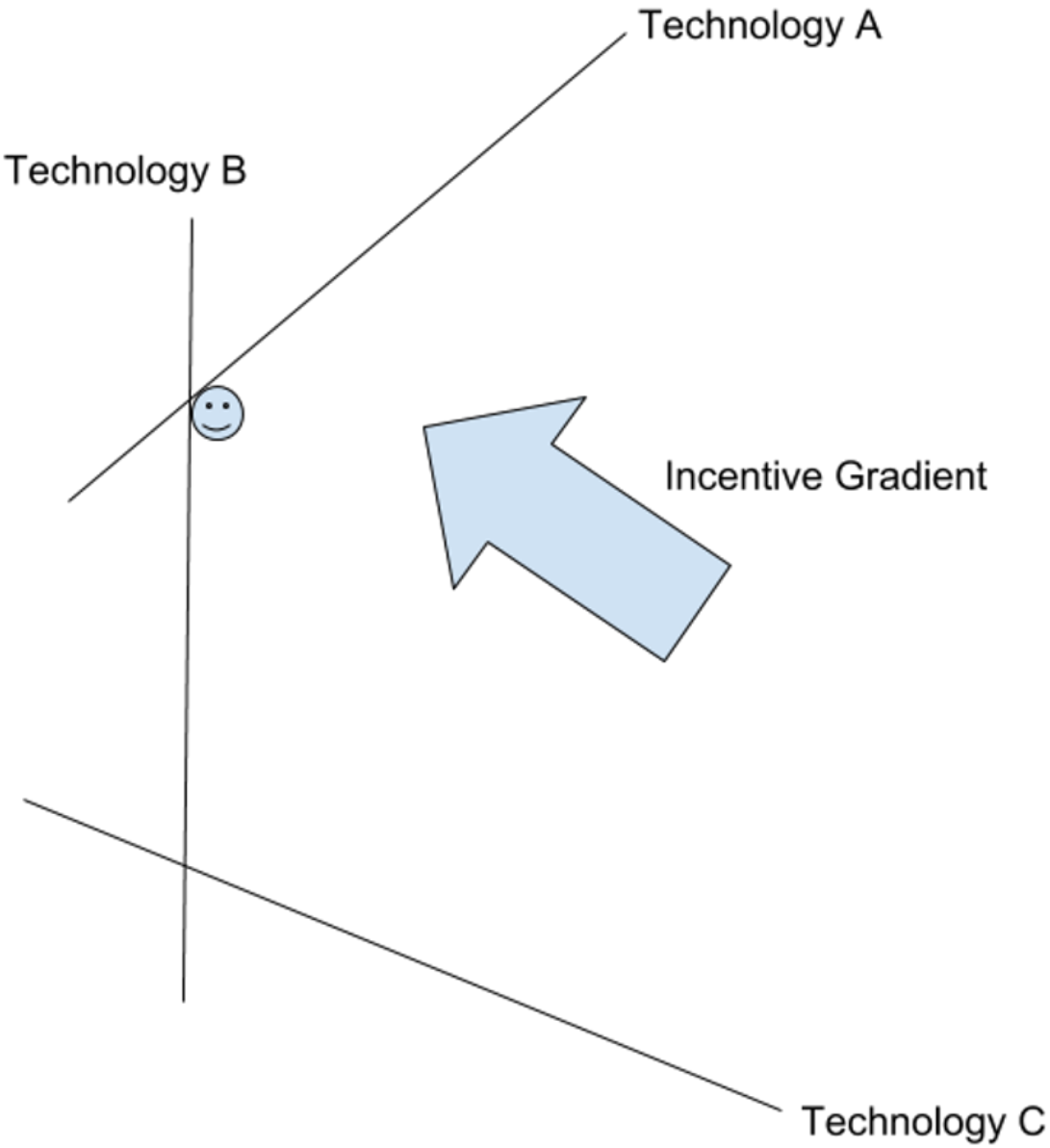
Once you go looking for it, there's a lot of technology shifts like this. Newcomen's steam engine (and Heron's, long before). Schwenteer's telegraph. Bushnell's submarine. Babbage and Lovelace had all the key ideas for the modern computer in the 1820's, but it wasn't until the 1890 census that somebody wanted to pay for such a thing. And of course, Moore's Law led to all sorts of ideas going from unprofitable to ubiquitous in the span of a decade or two.

In all these cases, the pattern is the same: the idea for an invention long predates the price shifts which make it marketable.

On the other hand, this isn't the case for *all* technological progress. There are some technologies for which demand preceded capability. After some insight or breakthrough made the technology possible, adoption followed rapidly. Consider the Wright brothers' flyer, or Edison's lightbulb. Both had badly inferior predecessors, which didn't really solve the problem: gliders and hot-air balloons for the Wright brothers, arc lights for Edison. Both built fast iteration platforms, tested a large possibility space, and eventually found a design which worked. And both saw rapid adoption once the invention was made.

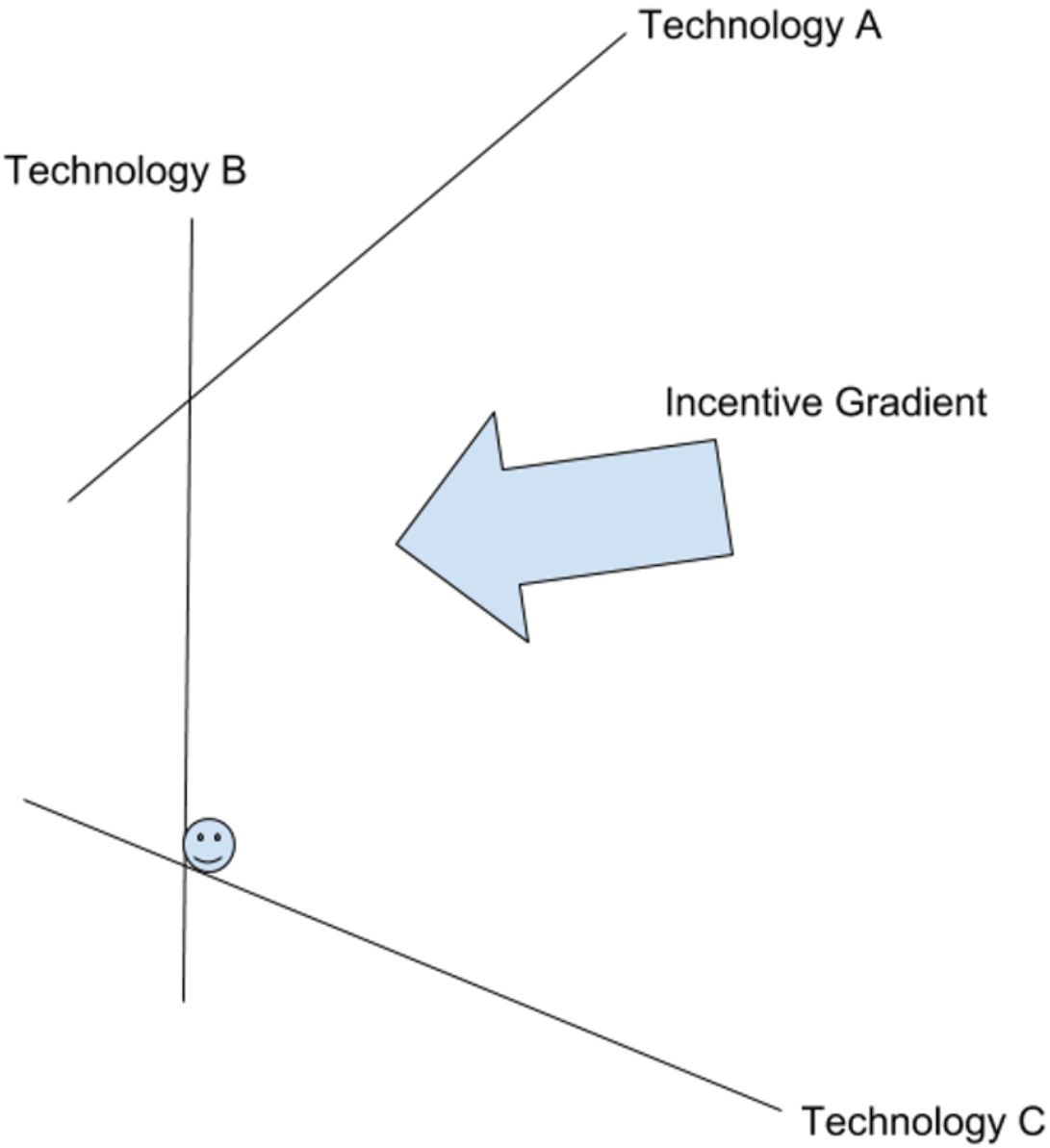
One notable feature of these breakthrough-type technologies: the economic incentive for flight or the lightbulb was in place long before the invention, so of course many people tried to solve the problems. Both Edison and the Wright brothers were preceded by many others who tried and failed.

Here's a simple model: technology determines the limits of what's possible, the constraints on economic activity. We can think of these constraints as planes in some high-dimensional space of economic production. Economic incentives push us as far as we can go along some direction, until we run in to one of these constraints—and the technology we use depends on what constraint we hit.



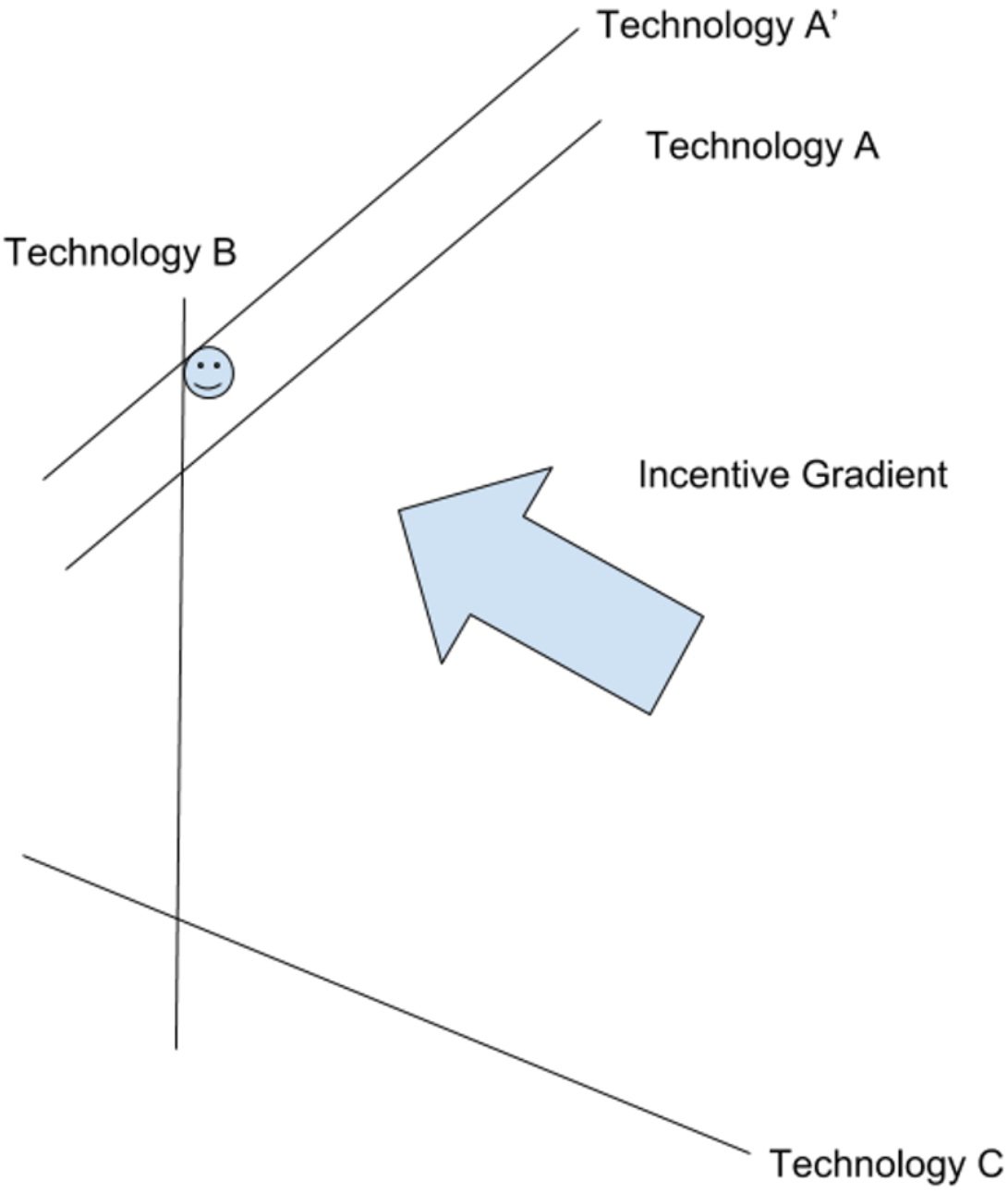
Following the incentive gradient in the diagram above, we end up at the smiley face—using a mix of technologies A and B. This point is insensitive to small changes in the incentive gradient—the prices can shift a bit one way or the other, shifting the incentive gradient slightly, and the smiley-face point will still be optimal.

However, if prices shift enough, then we can see a sudden change.



Once the incentive gradient moves “down” sufficiently, we suddenly jump from the A-B intersection being optimal to the B-C intersection being optimal. A new set of constraints kicks in; we switch from technology A to technology C. That’s the printing press: inventing C doesn’t matter until the prices shift.

On the other hand, we can also change technologies by relaxing a constraint. Suppose some new-and-improved version of technology A comes along:



Technology A' allows us to ignore the old A constraint, and move further along that direction. If we were using A before, then we'll definitely want to switch to A' right away. That's Edison's lightbulb.

In order for a technology to go from not-used to used, one of these two situations must hold: either the technology was unprofitable before and a price shift makes it profitable, or else it *was* profitable before, and many people tried to figure it out but couldn't. If you yourself want to market some kind of technology, then you should consider which of these two situations applies. Has a recent price shift made it profitable? Have you made some sort of breakthrough which others have tried and failed to find? If the answer to both of those questions is no, then the technology will probably remain unused. If the answer to at least one of those questions is yes, then you may be on to something.

Population Aging as an Impediment to Addressing Global Catastrophic Risks

[epistemic status: first Less Wrong post, developing hypothesis, seeking feedback and help fleshing out the hypothesis into something that could be researched and about which a discussion paper can be written. A comment/contribution to Eliezer Yudkowsky's "Cognitive biases potentially affecting judgment of global risks" in Bostrom & Cirkovic's "Global Catastrophic Risks" (2008)]

Most of the Global Catastrophic Risks we face in the 21st century, like anthropogenic climate change, comet and asteroid impacts, pandemics, and uncontrolled artificial intelligence, are high impact (affecting the majority or all of humanity), of terminal intensity (producing mass death, economic and social disruption, and in some cases potential human extinction), and are of highly uncertain probability [1]. This last factor is one major factor making it difficult to bring public attention and political will to bear on mitigating them. This is critical as all of our work/research on AI safety and other issues will be for naught if there is no understanding or will to implement them. Implementation may not require public involvement in some cases (AI safety may be manageable by consensus between AI researchers, for example) but others, like the detection of Earth orbit crossing asteroids and comets, may require significant expenditure to build detectors, etc.

My interest at present is in additional factors that make mustering political and public will even more difficult - given that these are hard problems to interest people in in the first place, what factors make that even more difficult? I believe that the aging of populations in the developed world may be a critical factor, progressively redirecting societal resources from long-term projects like advanced infrastructure, or foundational basic science research (which arguably AI Safety counts as), towards provision of health care and pensions.

Several factors make an aging developed world population a factor in blunting long-term planning:

(1) Older people (age 65+), across the developed world, vote more often than younger people

(2) Voters are more readily mobilized to vote to protect entitlements than to make investments for the future

(3) Older voters have access to, and are more aware of, entitlements than are younger people

(4) Expanding on (3), Benefits and entitlements are of particularly high salience to the aged because of their failure to save adequately for retirement. This trend has been ongoing and seems unlikely to be due to cognitive biases surrounding future planning.

(6) Long term investments, research, and other protections/mitigations against Global Catastrophic Risks will require a tradeoff with providing benefits to present people

(7) Older people have more present focus and less future focus than younger people (to the extent that younger people do - my anecdotal data is that most people interested in the far future of humanity are <50 years old, and a small subset of that

<50 year old population). Strangely, even people with grandchildren and great-grandchildren express limited interested in how their descendants will live and how safe their futures will be.

#6 is the point on which I am most uncertain (though I welcome questions and challenges that I should be more uncertain). Unless artificial intelligence and automation in the near term (15-30 years) provide really substantial economic benefits, enough that adequate Global Catastrophic Risk mitigation could be requisitioned without everyone noticing too much (and even then it may be a hard sell), it seems likely that future economic growth will be slower. Older workers, on average (my hunch says ...) are harder to retrain, and harder to motivate to retrain to take new positions, especially if the alternative is state-funded retirement. In a diminished economic future, one not as rich as it would have been with a more stable population pyramid, politics seems likely to focus on zero-sum games of robbing (young) Peter to pay (old) Paul, whether directly through higher taxation or indirectly by under-investing in the future.

Am I jumping ahead of the problem here? Do we not know enough about what it would take to address the different classes of Global Catastrophic and Existential Risk, or is there a reason to focus now on the factors that could prevent us from 'doing something about it'?

The Art of the Overbet

Previously: The Kelly Criterion

Last time I said, never go full Kelly.

In practice, I strongly agree with this. Either one should go *far over* full Kelly because the core Kelly assumptions have broken down and you want to throw the rules out the window, *or* you are trying to responsibly grow a bankroll and full Kelly betting on what you believe your edge to be would be massively overly aggressive.

This time we go over the second category with practical examples. What are situations in which one should engage in what *appears to be* massive over betting?

Four main scenarios: Not winning is catastrophic, losing is acceptable, and losing is impossible, losing is inevitable.

You Win or You Die

When you play the game of thrones, you win or you die. If playing it 'safe' with your resources means you don't win, then it means that you die. Either don't play, or don't hold back.

In Rounders (spoiler alert, movie is recommended), the star finds himself owing $15,000 to the Russian mob, and his best efforts could only get him $10,000. Without better options and unwilling to run for his life, he walks *into the poker club of the man he owes the money to,* and says "I owe you that money tomorrow, right? I've got $10,000. I'm looking for a game."

Famously, early in the company's history, the founder of UPS once found himself without the funds to make payroll. He knew that if he missed payroll, that would be the end of UPS. So he flew to Vegas with what funds he had, bet them all on black, won, made payroll, and now we have a UPS.

Magic players often select 'safe' decks instead of 'risky' decks. This is exactly backwards. If you flame out of a tournament and lose all rounds, you get nothing. If you win half of them, you still get nothing. If you win three quarters, you usually *still* get almost nothing relative to winning. Extra variance is great.

If you believe that the fate of everyone depends on crossing a threshold or hitting an impossibly precise target, no matter how bad the odds, you deploy everything you have and take your best shot.

There's a deadline. We don't have time for Kelly.

Win or Go Home

Losing not being so bad, or continuing to play not being so good, is effectively *very similar* to losing being catastrophic but not safely avoidable. In both cases, *surviving* is not a worthwhile goal.

A classic mistake is the gambler who forgets that they are using their bankroll to pay the rent. On forums I would often read stories of hard working sports gamblers with mid-five figure bankrolls. They make sure to make only ‘responsible’ wagers on sporting events, risking only 1-3% of their funds each time. When they have a good month, they could eat and make rent, but that ate up most of the profits.

What they continuously failed to realize was that this was not a sustainable situation. Being ‘responsible’ only ensured that, *even if they were good enough at picking winners,* they would never succeed due to their fixed costs. They were being ‘responsible’ with their sizing relative to their bankroll, but completely irresponsible when sizing relative to fixed costs of their time.

When I first started gambling on sports, I put aside a fixed bankroll I committed not to replenishing, but then acted what any Kelly-style formula would call completely irresponsible with that bankroll until after my first few double ups. As the operation became clearly worth my time, I scaled back and did more ‘responsible’ sizing, knowing that I wouldn’t be eaten alive by fixed costs.

Later, when exploring if it was worthwhile to return to wagering, I did a similar thing, struggled, and eventually lost everything I’d set aside. This was very good. It let me move on. To this day, we can never know for sure whether I had an edge during that second attempt – I suspect my luck was quite perverse – but I do know it wasn’t the best use of my time to find out.

Fail fast is a well-known principle for start-ups. I broke this rule once, with MetaMed, and it was an expensive mistake. Even if you have segregated your potential losses in dollar space, you need to contain them in time space, and emotional space, and social space.

There’s no deadline. But we don’t have time for Kelly.

Bet You Can’t Lose

You can’t lose everything if you can’t risk everything.

Most people’s resources are mostly not money, or even things at all, most of the time. When someone ‘loses everything’ we talk of them losing their friends, their family, their reputation, their health, their ability to work. And for good reason.

There is a scene in Defending Your Life where we flash back to the main character paying over $2,000, a third of all the money he has, to avoid a middle seat on an international flight. In context, this is to his credit, because it was ‘brave’ to risk spending so much. In a sense it was brave, in a more important general sense it was stupid, but in the most important sense he was spending a very small fraction of his resources, so the relevant questions were: Was this transaction worth it? No. Did this put the character at risk of having a liquidity crisis where not having any cash could be expensive? A little.

The best reason to do ‘safe’ things with money, and to maintain liquid savings, is to avoid paying the cost of needing liquidity and not having it, or paying the costs of avoiding scenarios where that liquidity might become necessary. This includes the ability to take advantage of opportunity. Avoiding extra costs of having to borrow money, which is expensive *in time* and *in emotional well-being and social capital,* not only in *money,* is by most people largely underestimated. Slack is vital.

There is also a *benefit* to having no cash. For some people, and in some families and cultures, one who has cash is expected to spend it, or inevitably will quickly spend it. Sometimes this is on one's self, sometimes on others, but the idea that one can have money and conserve it by spending only responsibly isn't a thing. The only socially acceptable way *to not spend money* is to *not have money.* Thus, there is a high effective 'tax' on savings.

In such situations, not only is risk not so bad, it can be actively great. If your risky bet pays off, you can have enough money for big purchases or even to escape your poverty trap. If it fails, you would have wasted the money anyway, and now you can conserve slash mooch off others. Thus, your goal is to find a way to translate cash, which one will inevitably lose, into inalienable property that can be protected, or skills and connections and relationships, or at least experiences one can remember.

This isn't just for poor people. *Start-up culture* works this way, too, and the only way to get things at reasonable prices, including labor, or to be able to raise money, is to clearly have spent everything you have and have no spare resources. This is a large portion of why start-ups are so stressful – you are not allowed to have any slack of any kind. I'm likely about to start a new company, and this is the thing that I dread most about the idea.

You Bet Your Life

This is all in contrast to the grand project we have been assigned, of 'saving for retirement.'

Older people saving for retirement in modern developed countries, or at least some of those saving for retirement, face a special situation. Unable to earn additional funds, and without families or friends they can count on, their survival or at least quality of life depends entirely upon their ability to save up money earlier in life to spend later.

If they run out of money, they'll still get some amount of government support, and perhaps some amount of familial support, but any remaining years are going to suck. If they die with money in the bank, that money can be passed on to others, but this is mostly a small benefit relative to not running out of cash to spend.

Compound this with large variance in remaining life span, and highly variant needs for health care and assistance during those remaining years, and you have an entire population pressured to furiously save every dollar they can. Any other use of one's resources is looked at as irresponsible.

This is *profoundly weird* and *profoundly messed up.*

It is handled, by most people and by those we trust to advise us, mindbogglingly badly.

That's true even in the cases where the problem definition mostly applies to one's situation, if one is willing to assume the world will continue mostly as it is, *and* one is unable to invest in other types of resources and expects to have little assistance from them.

It also leads us, as a society, to treat our savings as *the* savings, the retirement savings, and to apply the principles of that problem to all saving problems, and all risk

management problems. Young people take ‘risks’ and buy stocks, older people play it ‘safe’ and buy bonds.

If the world were a much more certain place, where we knew how long we would live, in what state of health, and how the economy and world would do, and everything else we might need money for, and how much money we needed to engage in various activities and consume various goods, we could *kind of* have a target number of dollars. There was an ad a few years back that was literally this. Each person had a giant seven-figure red number they would carry around from place to place, with an oddly exact number of dollars they ‘needed’ in order to be ‘able to’ retire. Implied was that less than this was failure and would be terrible, more than that would be success and no further funds required. Now you can rest.

Instead, at best we have a probabilistic distribution of how much utility one would be able to get from various amounts of capital, in various scenarios involving one’s life and how the outside world is working. Even in this simplified problem, how does one then ‘play it safe’? At best one can reduce variance from some ‘normal’ shocks like a decline in the stock market, while still being exposed to others, and likely not protect much at all from bigger risks. No matter what, the bulk of what you have will *probably* go to waste or be passed on to others, or else you are risking disaster. At a minimum, you’re ‘betting’ on what to do with and where to spend the rest of your life, and there you are very much all-in. Odd for someone who is about to die to even *try* to ‘play it safe.’

Introducing the AI Alignment Forum (FAQ)

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

After a few months of open beta, the AI Alignment Forum is ready to launch. It is a new website built by the team behind LessWrong 2.0, to help create a new hub for technical AI Alignment research and discussion. This is an in-progress FAQ about the new Forum.

What are the five most important highlights about the AI Alignment Forum in this FAQ?

- The vision for the forum is of **a single online hub** for alignment researchers to have conversations about **all ideas in the field**...
- ...while also **providing a better onboarding experience** for people getting involved with alignment research than exists currently.
- There are **three new sequences** focusing on some of the major approaches to alignment, which **will update daily for the coming 6-8 weeks**.
    - Embedded Agency, *written by Scott Garrabrant and Abram Demski of MIRI*
    - Iterated Amplification, *written and compiled by Paul Christiano of OpenAI*
    - Value Learning, *written and compiled by Rohin Shah of CHAI*
- For **non-members and future researchers, the place to interact with the content is LessWrong.com**, where all Forum content will be crossposted.
- The **site will continue to be improved in the long-term**, as the team comes to better understands the needs and goals of researchers.

What is the purpose of the AI Alignment Forum?

Our first priority is obviously to avert catastrophic outcomes from unaligned Artificial Intelligence. We think the best way to achieve this at the margin is to build an online-hub for AI Alignment research, which both allows the existing top researchers in the field to talk about cutting-edge ideas and approaches, as well as the onboarding of new researchers and contributors.

We think that to solve the AI Alignment problem, the field of AI Alignment research needs to be able to effectively coordinate a large number of researchers from a large number of organisations, with significantly different approaches. Two decades ago we might have invested heavily in the development of a conference or a journal, but with the onset of the internet, an online forum with its ability to do much faster and more comprehensive forms of peer-review seemed to us like a more promising way to help the field form a good set of standards and methodologies.

Who is the AI Alignment Forum for?

There exists an interconnected community of Alignment researchers in industry, academia, and elsewhere, who have spent many years thinking carefully about a variety of approaches to alignment. Such research receives institutional support from organisations including FHI, CHAI, DeepMind, OpenAI, MIRI, Open Philanthropy, and others. The Forum membership currently consists of researchers at these organisations and their respective collaborators.

The Forum is also intended to be a way to interact with and contribute to the cutting edge research for people not connected to these institutions either professionally or socially. There have been many such individuals on LessWrong, and that is the current best place for such people to start contributing, to be given feedback and skill-up in this domain.

There are about 50-100 members of the Forum. These folks will be able to post and comment on the Forum, and this group will not grow in size quickly.

Why do we need another website for alignment research?

There are many places online that host research on the alignment problem, such as the OpenAI blog, the DeepMind Safety Research blog, the Intelligent Agent Foundations Forum, AI-Alignment.com, and of course LessWrong.com.

But none of these spaces are set up to host discussion amongst the 50-100 people working in the field. And those that do host discussion have unclear assumptions about what's common knowledge.

What type of content is appropriate for this Forum?

As a rule-of-thumb, if a thought is something you'd bring up when talking to someone at a research workshop or a colleague in your lab, it's also a welcome comment or post here.

If you'd like a sense of what other Forum members are interested in, here's some quick data on what high-level content forum members are interested in seeing, taken from a survey we gave to invitees to the open beta (n = 34).

The responses were on a 1-5 scale, which represented "If I see 1 post per day, I want to see this type of content..." (1) Once per year, (2) Once per 3-4 months (3) Once per 1-2 months (4) Once per 1-2 weeks (5) A third of all posts that I see.

Here were the types of content asked about, and the mean response:

- New theory-oriented alignment research typical of MIRI or CHAI: **4.4 / 5**

- New ML-oriented alignment research typical of OpenAI or DeepMind's safety teams: **4.2 / 5**
- New formal or nearly-formal discussion of intellectually interesting topics that look questionably/ambiguously/peripherally alignment-related: **3.5 / 5**
- High-quality informal discussion of alignment research methodology and background assumptions, what's needed for progress on different agendas, why people are pursuing this or that agenda, etc: **4.1 / 5**
- Attempts to more clearly package/explain/summarise previously discussed alignment research: **3.7 / 5**
- New technical ideas that are clearly not alignment-related but are likely to be intellectually interesting to forum regulars: **2.2 / 5**
- High-quality informal discussion of very core background questions about advanced AI systems: **3.3 / 5**
- Typical AGI forecasting research/discussion that isn't obviously unusually relevant to AGI alignment work: **2.2 / 5**

Related data: After integrating over all 34 respondents' self-predictions, they predict 3.2 comments and 0.99 posts per day. We'll report on everyone's self-accuracy in a year ;)

What are the three new sequences I've been hearing about?

We have been coordinating with AI alignment researchers to create three new sequences of posts that we hope can serve as introductions to some of the most important core ideas in AI Alignment. The three new sequences will be:

- Embedded Agency, written by Scott Garrabrant and Abram Demski of MIRI
- Iterated Amplification, written and compiled by Paul Christiano of OpenAI
- Value Learning, written and compiled by Rohin Shah of CHAI

Over the next few weeks, **we will be releasing about one post per day from these sequences**, starting with the first post in the Embedded Agency sequence.

If you are interested in learning about AI alignment, you're very welcome to ask questions and discuss the content in the comment sections. And if you are already familiar with a lot of the core ideas, then we would greatly appreciate feedback on the sequences as we publish them. We hope that these sequences can be a major part of how new people get involved in AI alignment research, and so we care a lot about their quality and clarity.

In what way is it easier for potential future Alignment researchers to get involved?

Most scientific fields have to balance the need for high-context discussion with other specialists, and public discussion which allows the broader dissemination of new ideas, the onboarding of new members and the opportunity for new potential

researchers to prove themselves. We tried to design a system that still allows newcomers to participate and learn, while giving established researchers the space to have high-level discussions with other researchers.

To do that, we integrated the new AI Alignment Forum closely with the existing LessWrong platform, where you can find and comment on all content on the AI Alignment Forum on LessWrong, and your comments and posts can be moved to the AI Alignment Forum by mods for further engagement by the researchers. For details on the exact setup, see the question on that below.

We hope that this will result in a system in which cutting-edge research and discussion can happen, while new good ideas and participants can get noticed and rewarded for their contributions.

If you've been interested in doing alignment research, then we think one of the best ways to do that right now is to comment on AI Alignment Forum posts on LessWrong, and check out the new content we'll be rolling out.

What is the exact setup with content on LessWrong?

Here are the details:

- **Automatic Crossposting** - Any new post or comment on the new AI Alignment Forum is automatically cross-posted to LessWrong.com. Accounts are also shared between the two platforms.
- **Content Promotion** - Any comment or post on LessWrong can be promoted by members of the AI Alignment Forum from LessWrong to the AI Alignment Forum.
- **Separate Reputation –** The reputation systems for LessWrong and the AI Alignment Forum are separate. On LessWrong you can see two reputation scores: a primary karma score combining karma from both sites, and a secondary karma score specific to AI Alignment Forum members. On the AI Alignment Forum, you will just see their AI Alignment karma.
- **Content Ownership** - If a comment or post of yours is promoted to the AI Alignment Forum, you will continue to have full ownership of the content, and you'll be able to respond directly to all comments by members on your content.

The AI Alignment Forum survey (sent to all beta invitees) received 34 submissions. One question asked **whether the integration with LW would lead to the person contributing more or less to the AI Alignment Forum** (on a range from 0 to 6). The mean response was 3.7, the median was 3, and there was only one response below 3 (where 3 represented 'doesn't matter').

How do new members get added to the Forum?

There are about 50-100 members of the AI Alignment Forum, and while the number will grow, it will grow rarely and slowly.

We’re talking with the alignment researchers at CHAI, DeepMind, OpenAI, MIRI, and will be bringing on a moderator with invite-power from each of those organisations. They will naturally have a much better sense of the field and researchers in their orgs, than we the site designers. We’ll edit this post to include them once they’re confirmed.

On alignmentforum.org in the top right corner (after you created an account) is a small application form available. If you’re a regular contributor on LessWrong and want to point us to some of your best work, or if perhaps you’re a full-time researcher in an adjacent field and would like to participate in the Forum research discussion, you’re welcome to use that to let us know who you are and what research you have done.

Who is running this project?

The AI Alignment Forum development team consists of Oliver Habryka, Ben Pace, Raymond Arnold, and Jim Babcock. We're in conversation with alignment researchers from DeepMind, OpenAI, MIRI and CHAI to confirm moderators from those organisations.

We would like to thank BERI, EA Grants, Nick Beckstead, Matt Wage and Eric Rogstad for the support that lead to this Forum being built.

Can I use LaTeX?

Yes! You can use LaTeX in posts and comments with Cmd+4 / Ctrl+4.

Also, if you go into your user settings and switch to the markdown editor, you can just copy-paste LaTeX into a post/comment and it will render when you submit with no further work.

(Talk to us in intercom if you run into any problems.)

I have a different question.

Use the comment section below. Alternatively, use intercom (bottom right corner).

You're never wrong injecting complexity, but rarely you're right

I just want to put this idea in the form of a post, to gather your impressions. I think it's my main rationalist failure mode.

Recently in a Facebook group, some poster has proposed this synthesis of Harari's book '21 lessons':

> In the 21st century, three narratives were used to explain the past and predict the future: the fascist, the communist and the liberal narrative. During the century, the latter has prevailed, although in recent times it has started to crack, due to events like the election of Trump, the Brexit, and so on.

Then, the same user asked: what could be a new narrative that would help us in the future?

I was tempted to reply as I always do: criticize the simplification. I was about to write that the concept of narrative itself is a narrative, that Harari is seeing the past with the eyes of the present, but not necessarily this lens will help with navigating the future, that also a better concept would be that of a memeplex, which is less internally coherent than a story, and thus more complex to pinpoint.

Then a reflection occured to me: I always end up doing this, in almost all discussions I partecipate. People simplify too much and come to the wrong conclusion, they consider only the extremes of a spectrum, they use words as rigid classifiers and debate endlessly about them, they do not have internally coherent point of views, etc. I almost invariable end up 'winning' (i.e. appear wise) by injecting some complexity: usually in the form of a new parameter that was buried in the presuppositions.

Then, I was struck by another insight: it is too easy to win this way. From a mathematical point of view, a system with a bigger state space is more flexible than a smaller system, and so an 'optimization' that increases the space state is always correct. Is that possible? How probable it is that I've discovered a universal optimization of every human debate? I reasoned that it's very low, and indeed I think I've always failed to consider the downside: a more complex system is more difficult to use. When I add a parameter to the problem space, I'm multiplying the costs to use said idea in practice.

A classic example: was Martin Luther King a criminal? Some argue from the letter of the law, some argue from a moral point of view. If I 'win' the debate by injecting the concept of 'moral inertia' (speed at which a legal system reacts to an evolving moral landscape), then I implicitly add a parameter to every closed case, that could be potentially re-examined.

So I've decided I will refrain to add complexity, to formally but sterilely win debates, unless I have also a good 'reduction', that forgets some complexity in favor of being more useful. Doing so however is not trivial, it adds real work: this is due to the fallacy of gray, the fact that there are shades of gray doesn't mean that perfect middle gray is the correct answer.

Computerphile discusses MIRI's "Logical Induction" paper

This is a linkpost for https://www.youtube.com/watch?v=gDqkCxYYDGk

Preface to the sequence on value learning

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

This is a meta-post about the upcoming sequence on Value Learning that will start to be published this Thursday. This preface will also be revised significantly once the second half of the sequence is fully written.

Purpose of the sequence

The first part of this sequence will be about the tractability of ambitious value learning, which is the idea of inferring a utility function for an AI system to optimize based on observing human behavior. After a short break, we will (hopefully) continue with the second part, which will be about why we might want to think about techniques that infer human preferences, even if we assume we won't do ambitious value learning with such techniques.

The aim of this part of the sequence is to gather the current best public writings on the topic, and provide a unifying narrative that ties them into a cohesive whole. This makes the key ideas more discoverable and discussable, and provides a quick reference for existing researchers. It is meant to teach the ideas surrounding *one* specific approach to aligning advanced AI systems.

We'll explore the specification problem, in which we would like to define the behavior we want to see from an AI system. Ambitious value learning is one potential avenue of attack on the specification problem, that assumes a particular model of an AI system (maximizing expected utility) and a particular source of data (human behavior). We will then delve into conceptual work on ambitious value learning that has revealed obstructions to this approach. There will be pointers to current research that aims to circumvent these obstructions.

The second part of this sequence is currently being assembled, and this preface will be updated with details once it is ready.

The first half of this sequence takes you near the cutting edge of *conceptual* work on the *ambitious* value learning problem, with some pointers to work being done at this frontier. Based on the arguments in the sequence, I am confident that the obvious formulation of ambitious value learning has major, potentially insurmountable conceptual hurdles given the ways that AI systems work currently, but it may be possible to pose a different formulation that does not suffer from these issues, or to add hardcoded assumptions to the AI system to avoid impossibility results. If you try to disprove the arguments in the posts, or to create formalisms that sidestep the issues brought up, you may very well generate a new interesting direction of work that has not been considered before.

There is also a community of researchers working on inverse reinforcement learning without focusing on its application to ambitious value learning; this is out of the scope

of the first half of this sequence, even though such work may still be relevant to long term safety.

Requirements for the sequence

Understanding these posts will require at least a passing familiarity with the basic principles of machine learning (*not* deep learning), such as "the parameters of a model are chosen to maximize the log probability that the model assigns to the observed dataset". No other knowledge about value learning is required. If you do not have this background, I am not sure how easy it will be to grasp the points made; many of the points feel intuitive to me even without an ML background, but this could be because I no longer remember what it was like to not have ML intuitions.

There are many different subcultures interested in AI safety, and the posts I have chosen to include involve linguistic choices and assumptions from different places. I have tried to make this sequence understandable to all people who are interested and who understand the basic principles of ML, and so if something seems odd/confusing, please do let me know, either in the comments or via the PM system.

Learning from this sequence

When collating this sequence, I tried to pick content that makes the most important points simply and concisely. I recommend reading through each post carefully, taking the time to understand each paragraph. The posts range from informal arguments to formal theorems, but even for the formal theorems the formalization of the problem could be changed to invalidate the theorem. Learn from this however you best learn; my preferred method is to try and disprove the argument in the post until I feel like I understand what the post actually conveys.

While this sequence as it stands has no exercises, what it does have is a surrounding forum and community. Here are a few actions you can take to aid both your and others' understanding of the core concepts:

- Leave a comment with a concise summary of what you understand to be the post/paper's main point
- Leave a comment outlining a confusion you have with paper/post
- Respond to someone else's comment to help them understand it better

While I can't commit to responding to the majority of the comments, I am also excited to help readers understand the content, and please let me know if something I write is confusing.

Each post has a note at the top saying what the post covers and who should read it. You can read through these notes and decide whether they are important for you. That said, the posts are written and organized assuming that you have read prior posts in the sequence, and many points will not make sense if read out of order.

Reflections on Being 30

Epistemic Status: Personal

I haven't written a lot of personal stuff here recently, because I've been doing a lot more private contemplation, and been busy with life things. (Nonprofit and baby, among other things.) But I thought I might want to put out some thoughts about what growing in maturity means to me and what I've come to believe — since I still believe firmly in the blogging medium and the practice of transparency.

Prudence

There's a transition that a lot of people go through as they get older, that has to do with "practicality" or "prudence." They no longer want to do things that will predictably fail. They are no longer as willing to deal with *people* who will predictably fail at life. They are no longer as interested in ideas that can't stand up to practical tests.

I've noticed more of this spirit in myself as I get older, but I've always been somewhat ambivalent about it. I like interestingness. I want to avoid the natural tendency to stop exploring as I age. Meeting new people, learning new things, having new experiences, expanding my boundaries, are still important to me.

On the other hand, I've really enjoyed becoming more adept at the "practical" or "operational" side of things — schedules, housework, childcare, managing a small organization, etc. My identity up until now has been "talented mess" — so much so that I got an ADHD diagnosis quite by accident, and am now exploring the very different world of practicality and detail-orientation and organization. It's strange. It's very calm, and it's a satisfying challenge to keep up with things and bring more order to different parts of my life, and it's completely *non-narrative*. Life becomes a series of tasks, rather than a story. I'm continually marveling that this is how some people have been living all along.

Of course, the real reason for being more prudent in your thirties or as a parent is hard necessity. You have less energy, and more responsibilities, and so you have to be more cautious with resources and time. This isn't something I really want to spin as a good thing — it's scarcity, plain and simple.

You have to give up *something*, and the cheapest thing to give up is being a dumbass.

I *want* to have an "abundance mentality", to be generous and spendthrift with my time and energy; but sometimes I come up against irreducible scarcity.

A friend advised me last year to "have an ego." He meant it in Freud's sense of the "rational self-interest" part of the psyche. An ego is an *institution* you build around yourself, like the Republic of Sarah, or Sarah Incorporated. Your household, your career, your reputation, your health, all these structures around yourself that you build and maintain and use to interface with the world.

So I did that.

I do a lot of adjusting and updating on these structures; in a sense that's most of what I do all day long. Taking care of my work, my family and household, my physical body,

etc. Like a hermit crab, the little soft emotional creature that is me is hidden within all this prudence and structure. I notice it works better. I notice people like it better. But I'm a little melancholy about it.

Humanism

One value I still hold very firmly is something I call "humanism", or being "pro-human" or believing in the worth of the human spirit. I don't think that has to go away with age.

The *whole* human mind, which is a *general* intelligence, which can learn anything and create anything, is a beautiful thing and not to be destroyed.

This is in contrast to some people who become traditionalists or authoritarians when they hit the age where they realize they need prudence. The temptation is to believe "people just need to be kept under tight enough control that they can't do dumb shit, because the consequences of doing dumb shit are tragic."

The thing is, I don't think that controlling people actually is a feasible way to prevent tragedy.

A child prevented from making mistakes isn't a perfect child, but an underdeveloped child.

If you manage to control someone's behavior well enough to "keep them out of trouble", there's a good chance you've damaged their ability to problem-solve, and — I don't know how to say this any other way — *injured the sacred thing that makes them human*.

People who say "autonomy is a figment, some people need to be controlled for their own good" are sometimes the same people who do *actually really bad things to human beings, by dehumanizing them.*

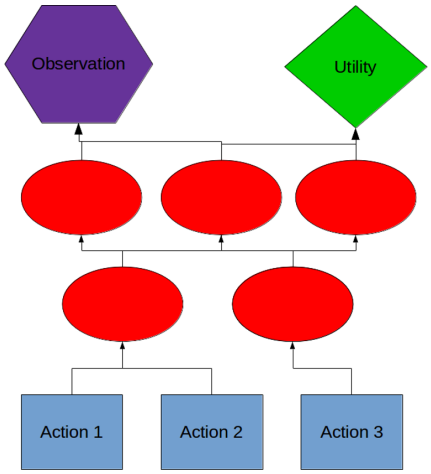
As I get less easily susceptible to opinions I hear, and more interested in the boring-but-true over the hot take, I become more humanist, not less so. It's not naive. It's actually looking at what people are, and noticing that they are a lot *more* complex and able than cynics give them credit for, that "people aren't all that special" or "some people aren't really people" is a brute's excuse.

You can totally be a mature person, or a parent, and still believe in humanism and autonomy. People have been doing it for hundreds of years.

Stop buttons and causal graphs

Edit: This work is known to be obsolete. The design will not make correct side bets on whether the button will be pressed.

Suppose that all problems relating to epistemic rationality have been solved. We create an AI with an epistemic module and an instrumental module. The epistemic module produces a causal graph like this one. (well a probability distribution over such graphs, but suppose only one has significant probability for now.)



Here blue rectangles are the actions the AI can take in the world, red ellipses represent real world events. The green diamond represents utility. Mathematically, these nodes are stochastic functions that take in several random variables inputs, and return an output random variable (RV). The rectangles are action nodes, so they just represent a function of the AI's action. Each node is the situation at a single instant, if the AI has multiple opportunities like this one at successive times, the graph needs to be repeated. The AI's instrumental rationality can work by $\text{argmax}_{Act}(E(\text{Utility}|A))$

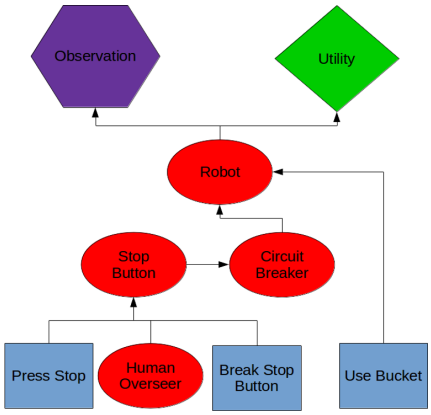
Now lets add a stop button, carrying on from the broomstick in this example. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/

The AI has 3 actions available. It can

1)Press its own stop button. After this, it can't do anything.

2)Break the stop button so the human can't press it, then fill the cauldron

3)Fill cauldron



$$U_{normal}(act) = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \end{cases}$$

If the AI has a utility U_{normal}, then it breaks the stop button, so the human can't stop it, then fill the bucket.

$$U_{suspend}(act) = \begin{cases} 1 & \text{if suspended} \\ 0 & \text{otherwise} \end{cases}$$

$$V = \begin{cases} U_{suspend} & \text{if Button Pressed} \\ U_{normal} & \text{if Button Not Pressed} \end{cases}$$

When we change the utility to V this AI presses the stop button. The referenced work proposed adding another term to the utility function.

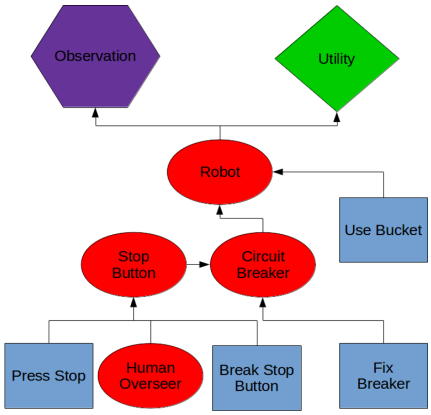
$$\theta = \max_{act} E[U_{normal} \mid act] - \max_{act} E[U_{suspend} \mid act]$$

$$V_{indifferent} = \begin{cases} U_{suspend} + \theta & \text{if Button Pressed} \\ U_{normal} & \text{if Button Not Pressed} \end{cases}$$

This is equivalent to the AI choosing $\text{argmax}_{Act}(E(U_{normal}|A, \text{Button Not Pressed}))$

This could be done in the causal graph by deleting all the inputs to the "Stop Button" node and setting that nodes value to definitely not pressed. This works, the AI chooses option 3) and fills the cauldron.

However suppose the situation was different. Here the "Circuit Breaker" is broken, and the AI has to fix it before the stop button will work. The AI now has a new option.



4)Fix Circuit Breaker, then use bucket.

The AI confidantly believes that the Stop Button will not be pressed, so it considers fixing the circuit breaker to be a waste of time.

Now lets cut the inputs to the “Stop Button” in the AI’s world model and set the node to be 50% pressed, 50% not pressed. Set the AI’s utility function to V. This AI believes that whether or not the button is pressed is an unalterable constant of nature. It can’t concieve of any action that would change the likelihood of the button being pressed. It will therefore not go out of its way to break the button, or to press it. If you put the button right in its path, it might press it accidentally. It will want to know if the button has been pressed, and will pay a small penalty to find out, or to fix the circuit breaker.

In utility function terms this is

$$V_{new}(A) = \begin{cases} U_{suspend}(A)/P(\text{Button On}|A) & \text{if Button Pressed} \\ U_{normal}(A)/P(\text{Button Off}|A) & \text{if Button Not Pressed} \end{cases}$$

So

$$E(V_{new}|A) = E(V_{new}|A, \text{Button Off})P(\text{Button Off}|A) + E(V_{new}|A, \text{Button On})P(\text{Button On}|A)$$

$$= E(U_{norm}|A, \text{Button Off}) + E(U_{suspend}|A, \text{Button On})$$

Before we were optimizing over all A for

$$E(V|A) = E(V|A, \text{Button Off})P(\text{Button Off}|A) + E(V|A, \text{Button On})P(\text{Button On}|A)$$

But by insisting that $P(\text{Button Off}|A) = P(\text{Button On}|A) = \frac{1}{2}$ we can have the AI choose

$$\text{argmax}_{Act}(\tfrac{1}{2}E(V|A, \text{Button Off}) + \tfrac{1}{2}E(V|A, \text{Button On}))$$

This gives both indifference to the buttons state, and gives the AI a value of information. Note that this AI will ignore a human hand moving towards the button, and switch off in response to seeing the button actually get pressed. The AI has no strong reason not to press the button. So this stops the AI making a serious effort to press the button, but if you leave its button on the floor, it won't avoid standing on it.

Does anyone think a mental intuition of optimization power flowing through causal networks from the actions to the utility function is useful?

Have I just solved the problem of utility switching, or is there a subtle failure mode I haven't thought of yet?

Mark Eichenlaub: How to develop scientific intuition

Recently on the CFAR alumni mailing list, someone asked a question about how to develop scientific intuition. In response, Mark Eichenlaub posted an excellent and extensive answer, which was so good that I asked for permission to repost it in public. He graciously gave permission, so I've reproduced his message below. (He otherwise retains the rights to this, meaning that the standard CC license on my blog doesn't apply to this post.)

From: **Mark Eichenlaub**
Date: Tue, Oct 23, 2018 at 9:34 AM
Subject: Re: [CFAR Alumni] Suggestions for developing scientific intuition

Sorry for the length, I recently finished a PhD on this topic. (After I wrote the answer kerspoon linked, I went to grad school to study the topic.) This is specifically about solving physics problems but hopefully speaks to intuition a bit more broadly in places.

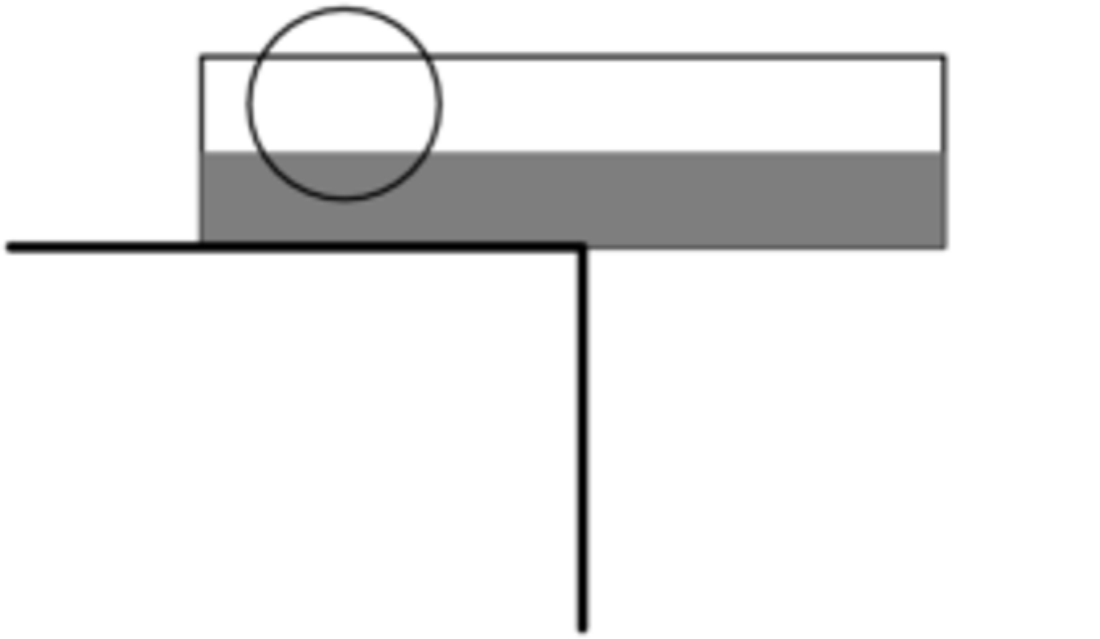
I mostly think of intuition as the ability to quickly coordinate a large number of small heuristics. We know lots of small facts and patterns, and intuition is about matching the relevant ones onto the current situation. The little heuristics are often pretty local and small in scope.

For example, the other day I heard this physics problem:

You set up a trough with water in it. You hang just barely less than half of the trough off the edge of a table, so that it balances, but even a small force at the far end would make it tip over.

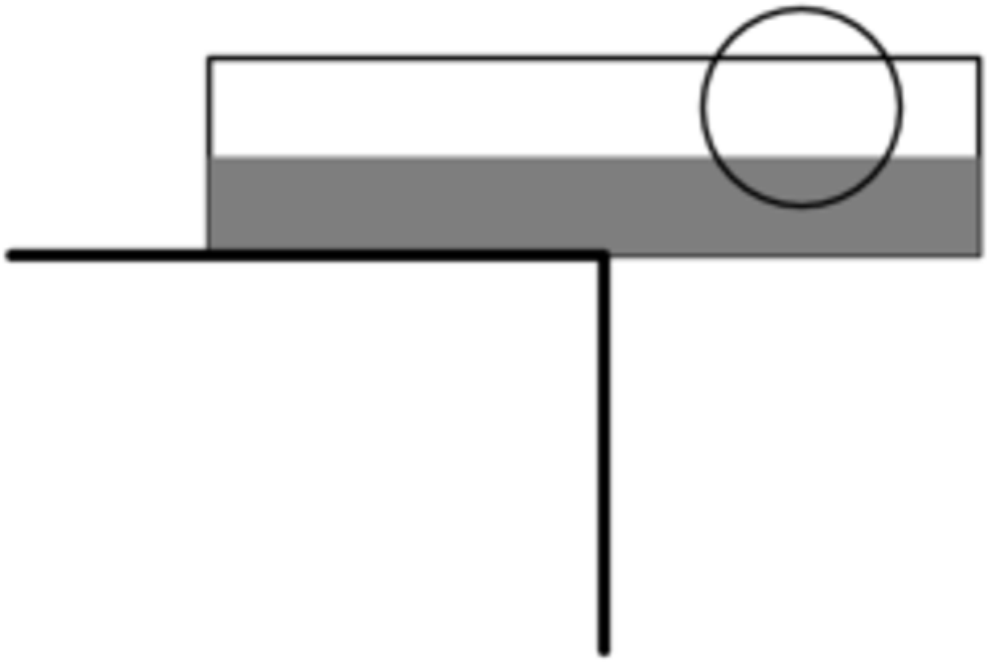
trough is barely balanced

You put a boat in the trough at the end over the table. The trough remains balanced.



add a boat to the left, trough still balanced

Then you slowly push the boat down to the other end of the trough, so that’s it’s in the part of the trough that hangs out from the table. What happens? (I.E. does the trough tip over?)



move boat to the right. What happens?

The answer is (rot13) Gur gebhtu qbrf abg gvc; vg erznvaf onynaprq (nf ybat nf gur zbirzrag bs gur obng vf fhssvpvragyl fybj fb gung rirelguvat erznvaf va rdhvyvoevhz).

I knew this “intuitively”, by which I mean I got it within a second or so of understanding the question, and without putting in conscious effort to thinking about it. (I wasn’t certain I was right until I had consciously thought it out, but I was reasonably confident within a second, and my intuition bore out.) I don’t think this was due to some sort of general intuition about problem solving, science, physics, mechanics, or even floating. It felt like I could solve the problem intuitively specifically because I had seen sufficiently-similar things that led me to the specific heuristic “a floating object spreads its weight out evenly over the bottom of the container it’s floating in.” Then I think of “having intuition” in physics as having maybe a thousand little rules like that and knowing when to call on which one.

For this particular heuristic, there is a classic problem asking what happens to the water level in a lake if you are in a boat with a rock, and you throw the rock into the water and it sinks to the bottom. One solution to that problem is that when the rock is on the bottom of the lake, it exerts more force on that part of the bottom of the lake than is exerted at other places. By contrast, when the rock is still in the boat, the only thing touching the bottom of the lake is water, and the water pressure is the same everywhere, so the weight of the rock is distributed evenly across the entire lake. The total force on the bottom of the lake doesn’t change between the two scenarios (because gravity pulls on everything just as hard either way), when the rock is sitting on the bottom of the lake and the force on the bottom of the lake is higher under the rock, it must be lower everywhere else to compensate. The pressure everywhere else is $\rho g h$, so if that goes down, the level of the lake goes down. Conclusion: when you throw the rock overboard, the level of the lake goes down a bit. When I thought about

that problem, I presumably built the “weight distributed evenly” heuristic. All I had to do was quickly apply it to the trough problem to solve that one as well.

And if someone else also had a background in physics but didn’t find the trough problem easy, it’s probably because they simply hadn’t happened to think about the boat problem, or some other similar problems, in the right way, and hadn’t come away with the heuristic about the weight of floating things being spread out evenly.

To me, this picture of intuition as small heuristics doesn’t look good for the idea of developing powerful intuition. The “weight gets spread out by floating” heuristic is not likely to transfer to much else. I’ve used it for two physics problems about floating things and, as far as I know, nothing else.

You can probably think of lots of similar heuristics. For example, “conservation of expected evidence“. You might catch a mistake in someone’s reasoning, or an error in a long probability calculation you made, if you happen to notice that the argument or calculation violates conservation of expected evidence. The nice thing about this is that it can happen almost automatically. You don’t have to stop after every calculation or argument and think, “does this break conservation of expected evidence?”. Instead, you wind up learning some sorts of triggers that you associate with the principle that prime it in your mind, and then, if it becomes relevant to the argument, you notice that and cite the principle.

In this picture, building intuition is about learning a large number of these heuristics, along with their triggers.

However, while the individual small heuristics are often the easiest things to point to in an intuitive solution to a problem, I do think there are more general, and therefore more transferrable parts of intuition as well. I imagine that the paragraph I wrote explaining the solution to the boat problem will be largely incomprehensible to someone who hasn’t studied physics. That’s partially because it uses concepts they won’t have a rigorous understanding of (e.g. pressure), that it tacitly uses small heuristics it didn’t explain (e.g. that the reason the pressure is the same along the bottom of the lake is that if it weren’t, there would be horizontal forces that push the water around until the pressure did equalize in this way), partially that it made simplifications that it didn’t state and it might not be clear are justified (e.g. that the bottom of the lake is flat). More importantly, it relies on a general framework of Newtonian mechanics. For example, there are a number of tacit applications of Newton’s laws in the argument. For example, I stated that the total force on the bottom of the lake is the same whether the rock is resting on the bottom or floating in the boat “because gravity pulls on everything just as hard either way”, but these aren’t directly connected concepts. Gravity pulls the system (boat + water + rock) down just as hard no matter where the rock is. That system is not accelerating, so by Newton’s second law, the bottom of the lake pushes up on that system just as hard in each scenario. And by Newton’s third law, the system pushes down on the bottom of the lake just as hard in each scenario. So understanding the argument involves some fairly general heuristics such as “apply Newton’s second law to an object in equilibrium to show that two forces on it have equal magnitude” – a heuristic I’ve used hundreds of times, and “decide what objects to define as part of a system fluidly as you go through a problem” (in this case, switching from thinking about the rock as a system to thinking about rock+boat+water as a single system) – a skill I’ve used hundreds to thousands of times across all of physics. (My job is to teach high schoolers to be really good at solving problems like this, so I spend way more time on it than most people, so applying a heuristic specific to solving introductory physics problems in a thousand independent instances is realistic for me.)

Then there may be more meta-level skills and heuristics that you develop in solving problems. These could be things like valuing non-calculation solutions, or believing that persevering on a tough problem is worthwhile. It's also important that intuition isn't just about having lots of little heuristics. It's about organizing them and calling the right one up at the right time. You'll have to ask yourself the right sorts of questions to prompt yourself to find the right heuristics, and that's probably a pretty general skill.

There is a fair amount of research on trying to understand what all these little heuristics are and how to develop them, but I'm mostly familiar with the research in physics.

In the Quora answer kerspoon linked, I cited George Lakoff, and I still that he's a good source for understanding how we go about taking primitive sorts of concepts (e.g. "up" and "down") and using and adapting them, via partial metaphor, to understanding more abstract things. For a specific example that's well-argued, see:

Wittmann, Michael C., and Katrina E. Black. "Mathematical actions as procedural resources: An example from the separation of variables." Physical Review Special Topics-Physics Education Research 11.2 (2015): 020114.

They argue that students understand the arithmetic action "separation of variables" via analogy to their physical understanding of taking things and physically moving them around. However, I think Wittman and Black's work is incomplete. For example, it doesn't explain why students using the motion analogy for separation of variables do it correctly – they could just as well use motion to encode algebraically-invalid rules. Also, they don't explain how the analogy develops. They just catalog that it exists.

A foundational work in trying to understand the components of physical intuition is:

DiSessa, Andrea A. "Toward an epistemology of physics." Cognition and instruction 10.2-3 (1993): 105-225.

This work establishes "phenomenological primitives"; little core heuristics such as "near is more", which are templates for physical reasoning. Drawing from these templates, we might conclude that the nearer you are to a speaker, the louder the sound, or that the nearer you are to the sun, the hotter it will be (and therefore that summer is hot because the Earth is nearer the sun – a false but common and reasonable belief).

That's a long and somewhat-obscure paper. I really like his student's work

Sherin, Bruce L. "How students understand physics equations." Cognition and instruction 19.4 (2001): 479-541.

Like Disessa, Sherin builds his own framework for what intuition is. His scope is more limited though, focusing solely on building and interpreting certain types of equations in a manner that combines "intuitive" physical ideas and mathematical templates. He spells this out in detail more in the paper, and it's incredibly clear and well-argued. Probably my favorite paper in the field.

A more general reference that's much more accessible than Disessa and more general an overview of cognition in physics than Sherin is
"How Should We Think About How Our Students Think" by my advisor, Joe Redish http://media.physics.harvard.edu/video/?id=COLLOQ_REDISH_093013 (video) https://arxiv.org/abs/1308.3911 (paper).

The actual process of building new heuristics is also studied, but over all I don't think we know all that much. See my friend Ben's paper

Dreyfus, Benjamin W., Ayush Gupta, and Edward F. Redish. "Applying conceptual blending to model coordinated use of multiple ontological metaphors." International Journal of Science Education 37.5-6 (2015): 812-838.

for an example of theory-building around how we create new intuitions. He calls on a framework from cognitive science called "conceptual blending" that is rather formal, but I think pretty entertaining to read.

A relevant search terms in the education literature:

"conceptual change"

but I find a lot of this literature to be hard-to-follow and not always a productive use of time to read.

On the applied side, I think the state of the art in evidence-backed approaches to building intuition, at least in physics, is modeling instruction. I'm not sure what the best introduction to modeling instruction is. They have a website that seems okay. Eric Brewe writes on it and he's usually very good. The basic idea is to have students collaboratively participate in the building of the theories of physics they're using (in a specific way, with guidance and direction from a trained instructor), which gets them to think about the "whys" involved with a particular theory or model in a way they usually wouldn't.

I have written some about why I think things like checking the extreme cases of a formula are powerful intuition-building tools. A preprint is available here: https://arxiv.org/pdf/1804.01639.pdf

However, I think it's dangerous to have rules like "always check the dimensions of your answer", "always check the extreme cases of a formula", or even "always check that the numbers come out reasonable." The reason is that having these things as procedures tends to encourage students to follow them by rote. A large part of the cognitive work involved isn't in checking the extreme cases or the dimensions, but in realizing that in this particular situation, that would be a good thing to do. If you're doing it only because an external prompt is telling you to, you aren't building the appropriate meta-level habits.
See https://www.tandfonline.com/doi/abs/10.1080/09500693.2017.1308037 for an example of this effect.

See papers on "metarepresentation" by Disessa and/or Sherin for another example of generalizable skills related to intuition and problem solving.

Unfortunately, I don't think writing books well or writing courses of individual study is something we know much about. I don't know anyone who has a significant grant for that; the most I've ever seen on it is a poster here or there at a conference. Generally, grants are awarded for improving high school and college courses, or for professional development programs, supporting department or institution level changes at schools, etc. So adults who just want to learn on their own are not really served much by the research on the area. If you're an adult who wants to self-study theoretical physics with an eye towards intuition, I recommend Leonard Susskind's series of courses "The Theoretical Minimum" (the first three courses exist as books, the rest only as video

lectures). He approaches mathematical topics with what I find an intuitive approach in most cases. Of course the Feynman lectures on physics are also very good.

I'll be building an introduction to physics course at Art of Problem Solving, starting work sometime this winter. It might be available in the spring, although students will mostly be middle and high school students (but anyone is welcome to take our courses). I currently teach an advanced physics problem-solving course at AoPS called "PhysicsWOOT". I try to support intuition-building practices there, but the main aim is in training these many small heuristics which students need to solve contest problems.

There should be something like modeling instruction for adult independent learners, but I don't know of it.

Alignment Newsletter #27

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

Dan Hendrycks has now joined, and will likely write summaries primarily on adversarial examples and robustness. As with Richard, his summaries are marked as such; I'm reviewing some of them now but expect to review less over time.

Highlights

80K podcast with Paul Christiano *(Paul Christiano and Rob Wiblin)*: This is a mammoth 4-hour interview that covers a lot of ground. I'll try to state the main points without the supporting arguments in roughly chronological order, listen to the podcast for more.

- The problem of AI safety is that we don't know how to build AI that does what we want it to do. It arises primarily because each actor faces a tradeoff between AI systems being maximally effective at its task, and being robustly beneficial.

- AI safety has had much more attention in the last few years.

- Everyone agrees that we don't know how to build AI that does what we want, but disagrees on how hard the problem is, or how it should be framed.

- The best arguments against working on alignment are opportunity cost (eg. working on biosecurity instead) and that the problem might be very easy or impossible, but even then it seems like work would be valuable for getting information about how hard the problem actually is.

- It's not very important for the best AI safety team to work with the best ML team for the purpose of pursuing alignment research, but it is important for actually building powerful aligned AI.

- The variance in outcomes from AGI come primarily from uncertainty in how hard the technical problem is, how people behave about AGI, and then how good we are at technical safety research. The last one is easiest to push on.

- It seems useful to build organizations that can make commitments that are credible to outsiders. This would allow the top AI actors to jointly commit that they meet a particular bar for safety, though this would also require monitoring and enforcing to be effective, which is hard to do without leaking information.

- We should expect slow takeoff, as Paul defines it. (I'm ignoring a lot of detail here.)

- We should focus on short timelines because we have more leverage over them, but the analogous argument for focusing on fast takeoff is not as compelling.

- Paul places 15% probability on human labor being obselete in 10 years, and 35% on 20 years, but doesn't think he has done enough analysis that people should defer to him.

- Comparing current AI systems to humans seems like the wrong way to measure progress in AI. Instead, we should consider what we'd be able to do now *if* AI becomes comparable to humans in 10-20 years, and compare to that.

- We can decompose alignment into the problem of training an AI given a smarter overseer, and the problem of creating a sufficiently smart overseer. These roughly correspond to distillation and amplification respectively in IDA. (There's more discussion of IDA, but it should be pretty familiar to people who have engaged with IDA before.) Reactions fall into three camps: a) IDA is hopelessly difficult, b) IDA is focusing on far-away problems that will be easy by the time they are relevant, and c) optimistic about IDA.

- Very few people think about how to solve the full problem, that is, solve alignment in the limit of arbitrarily intelligent AI. MIRI doesn't think about the question because it seems obviously doomed to them, while the broader ML community wants to wait until we know how to build the system. The other approaches are debate (AN #5), which is very related to IDA, and inverse reinforcement learning (IRL). However, there are key problems with IRL, and research hasn't shed much light on the core of the problem.

- AI safety via debate also shares the insight of IDA that we can use AI to help us define a better training signal for AI. (There's discussion of how debate works, that again should be familiar to anyone who has engaged with it before.) The biggest difficulty is whether human judges are actually capable of judging debates on truthfulness and usefulness, as opposed to eg. persuasiveness.

- There are three main categories of work to be done on IDA and debate -- engineering work to actually build systems, philosophical work to determine whether we would be happy with the output of IDA or debate, and ML research that allows us to try out IDA or debate with current ML techniques.

- We should focus on prosaic AI, that is, powerful AI built out of current techniques (so no unknown unknowns). This is easier to work on since it is very concrete, and even if AGI requires new techniques, it will probably still use current ones, and so work aligning current techniques should transfer. In addition, if current techniques go further than expected, it would catch people by surprise, which makes this case more important.

- With sufficient computation, current ML techniques can produce general intelligence, because evolution did so, and current ML looks a lot like evolution.

- The biggest crux between Paul and MIRI is whether prosaic AI can be aligned.

- One problem that MIRI thinks is irresolvable is the problem of inner optimizers, where even if you optimize a perfectly constructed objective function that captures what we want, you may create a consequentialist that has good behavior in training environments but arbitrarily bad behavior in test environments. However, we could try to solve this through techniques like adversarial training.

- The other problem is that constructing a good objective is incredibly difficult, and existing schemes are hiding the magic somewhere (for example, in IDA, it would be hidden in the amplification step).

- Research of the kind that MIRI does will probably be useful for answering the philosophical questions around IDA and debate.

- Ought's Factored Cognition (AN #12) project is very related to IDA.

- Besides learning ML, and careers in strategy and policy, Paul is excited for people to start careers studying problems around IDA from a CS lens, a philosophical lens, or a psychological lens (in the sense of studying how humans decompose problems for IDA, or how they judge debates).

- Computer security problems that are about attacking AI (rather than defending against attacks in a world with AI) are often very related to long term AI alignment.

- It is important for safety researchers to be respectful of ML researchers, since they are justifiably defensive given the high levels of external interest in safety that's off-base.

- EAs often incorrectly think in terms of a system that has been given a goal to optimize.

- Probably the most important question in moral philosophy is what kinds of unaligned AI would be morally valuable, and how they compare to the scenario where we build an aligned AI.

- Among super weird backup plans, we could build unaligned AI that is in expectation as valuable as aligned AI, which allows us to sidestep AI risk. For example, we could simulate other civilizations that evolution would produce, and hand over control of the world to a civilization that would have done the same thing if our places were swapped. From behind a veil of ignorance of "civilizations evolution could have produced", or from a multiverse perspective, this has the same expected value as building an *aligned* AI (modulo the resource cost of simulations), allowing us to sidestep AI risk.

- We might face an issue where society worries about being bigoted towards AI and so gives them rights and autonomy, instead of focusing on the more important question of whether their values or goals align with ours.

Rohin's opinion: This is *definitely* worth reading or listening if you haven't engaged much with Paul's work before, it will probably be my go-to reference to introduce someone to the approach. Even if you have, this podcast will probably help tie things together in a unified whole (at least, it felt that way to me). A lot of the specific things mentioned have been in the newsletter before, if you want to dig up my opinions on them.

Technical AI alignment

Technical agendas and prioritization

80K podcast with Paul Christiano *(Paul Christiano and Rob Wiblin)*: Summarized in the highlights!

The Rocket Alignment Problem *(Eliezer Yudkowsky)* (summarized by Richard): Eliezer explains the motivations behind MIRI's work using an analogy between aligning AI and designing a rocket that can get to the moon. He portrays our current theoretical understanding of intelligence as having massive conceptual holes; MIRI is trying to clarify these fundamental confusions. Although there's not yet any clear path from

these sorts of advances to building an aligned AI, Eliezer estimates our chances of success without them as basically 0%: it's like somebody who doesn't understand calculus building a rocket with the intention of manually steering it on the way up.

Richard's opinion: I think it's important to take this post as an explication of MIRI's mindset, not as an argument for that mindset. In the former role, it's excellent: the analogy is a fitting one in many ways. It's worth noting, though, that the idea of only having one shot at success seems like an important component, but isn't made explicit. Also, it'd be nice to have more clarity about the "approximately 0% chance of success" without advances in agent foundations - maybe that credence is justified under a specific model of what's needed for AI alignment, but does it take into account model uncertainty?

Agent foundations

EDT solves 5 and 10 with conditional oracles *(jessicata)*

A Rationality Condition for CDT Is That It Equal EDT (Part 1) *(Abram Demski)*

Ramsey and Joyce on deliberation and prediction *(Yang Liu et al)*

Learning human intent

Few-Shot Goal Inference for Visuomotor Learning and Planning *(Annie Xie et al)*

Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow *(Xue Bin Peng et al)*

Video Imitation GAN: Learning control policies by imitating raw videos using generative adversarial reward estimation *(Subhajit Chaudhury et al)*

Handling groups of agents

M^3RL: Mind-aware Multi-agent Management Reinforcement Learning *(Tianmin Shu et al)*

Interpretability

Stakeholders in Explainable AI *(Alun Preece et al)* (summarized by Richard): There are at least four groups for whom "explainable" AI is relevant: developers (who want AI to be easier to work with), theorists (who want to understand fundamental properties of AI), ethicists (who want AI to behave well) and users (who want AI to be useful). This has complicated work on explanability/interpretability: the first two groups focus on understanding how a system functions internally (described in this paper as "verification"), while the latter two focus on understanding what the system does ("validation"). The authors propose an alternative framing of interpretability, based on known knowns, unknown knowns, etc.

Training Machine Learning Models by Regularizing their Explanations *(Andrew Slavin Ross)*

Adversarial examples

Towards Deep Learning Models Resistant to Adversarial Attacks *(Aleksander Madry et al)* (summarized by Dan H): Madry et al.'s paper is a seminal work which shows that some neural networks can attain more adversarial robustness with a well-designed adversarial training procedure. They train networks on adversarial examples generated by several iterations of projected gradient descent rather than examples generated in one step (FGSM). Another crucial component is that they add slight noise to a clean example before generating a corresponding adversarial example. When trained long enough, some networks will attain more L-infinity adversarial robustness.

Dan H's opinion: What's notable is that this paper has survived third-party security analysis, so this is a solid contribution. This contribution is limited by the fact that its improvements are limited to L-infinity adversarial perturbations on small images, as follow-up work has shown.

Towards the first adversarially robust neural network model on MNIST *(Lukas Schott, Jonas Rauber et al)* (summarized by Dan H): This recent pre-print claims to make MNIST classifiers more adversarially robust to different L-p perturbations. The basic building block in their approach is a variational autoencoder, one for each MNIST class. Each variational autoencoder computes the likelihood of the input sample, and this information is used for classification. They also demonstrate that binarizing MNIST images can serve as strong defense against some perturbations. They evaluate against strong attacks and not just the fast gradient sign method.

Dan H's opinion: This paper has generated considerable excitement among my peers. Yet inference time with this approach is approximately 100,000 times that of normal inference (10^4 samples per VAE * 10 VAEs). Also unusual is that the L-infinity "latent descent attack" result is missing. It is not clear why training a single VAE does not work. Also, could results improve by adversarially training the VAEs? As with all defense papers, it is prudent to wait for third-party reimplementations and analysis, but the range of attacks they consider is certainly thorough.

Robustness

Bayesian Policy Optimization for Model Uncertainty *(Gilwoo Lee et al)*

Reinforcement Learning with Perturbed Rewards *(Jingkang Wang et al)*

Miscellaneous (Alignment)

Existential Risk, Creativity & Well-Adapted Science *(Adrian Currie)*: From a brief skim, it seems like this paper defines "creativity" in scientific research, and argues that existential risk research needs to be creative. Research is creative if it is composed of "hot" searches, where we jump large distances from one proposed solution to another, with broad differences between these solutions, as opposed to "cold" searches, in which we primarily make incremental improvements, looking over a small set of solutions clustered in the neighborhood of existing solutions. The paper argues that research on existential risk needs to be creative, because many aspects of such research make it hard to analyze in a traditional way -- we can't perform controlled experiments of extinction, nor of the extreme circumstances under which it is likely; there are many interdependent parts that affect each other (since existential risks typically involve effects on many aspects of society), and there is likely to be a huge amount of uncertainty due to lack of evidence. As a result, we want to change the norms around existential risk research from the standard academic norms, which

generally incentivize conservatism and "cold" searches. Table 1 provides a list of properties of academia that lead to conservatism, and asks that future work think about how we could mitigate these.

Rohin's opinion: While I'm not sure I agree with the reasons in this paper, I do think we need creativity and "hot" searches in technical AI safety, simply based on the level of confusion and uncertainty that we (or at least I) have currently. The properties in Table 1 seem particularly good as an initial list of things to target if we want to make creative research more likely.

AI strategy and policy

Countering Superintelligence Misinformation *(Seth Baum)* (summarized by Richard): Two ways to have better discussions about superintelligence are correcting misconceptions, and preventing misinformation from being spread in the first place. The latter might be achieved by educating prominent voices, creating reputational costs to misinformers (both individuals and companies), focusing media attention, etc. Research suggests the former is very difficult; strategies include addressing pre-existing motivations for believing misinformation and using advance warnings to 'inoculate' people against false claims.

Richard's opinion: I'm glad to see this systematic exploration of an issue that the AI safety community has consistently had to grapple with. I would have liked to see a more nuanced definition of misinformation than "information that is already clearly false", since it's not always obvious what qualifies as clearly false, and since there are many varieties of misinformation.

Prerequisities: Superintelligence Skepticism as a Political Tool

Other progress in AI

Exploration

The Dreaming Variational Autoencoder for Reinforcement Learning Environments *(Per-Arne Andersen et al)*

EMI: Exploration with Mutual Information Maximizing State and Action Embeddings *(Hyoungseok Kim, Jaekyeom Kim et al)*

Reinforcement learning

Near-Optimal Representation Learning for Hierarchical Reinforcement Learning *(Ofir Nachum et al)* (summarized by Richard): This paper discusses the use of learned representations in hierarchical RL. In the setting where a higher-level policy chooses goals which lower-level policies are rewarded for reaching, how bad is it when the goal representation isn't able to express all possible states? The authors define a metric for a representation's lossiness based on how close to optimal the policies which can be learned using that representation are, and prove that using a certain objective function, representations with bounded lossiness can be learned. They note a similarity between this objective function and those of mutual information estimators.

The authors test their learner on the MuJoCo Ant Maze environment, achieving compelling results.

Richard's opinion: This is a fairly mathematical paper and I didn't entirely follow the proofs, so I'm not sure how dependent they are on the particular choice of objective function. However, the empirical results using that objective seem very impressive, and significantly outperform alternative methods of learning representations.

Introducing Holodeck *(joshgreaves32)*

Generalization and Regularization in DQN *(Jesse Farebrother et al)*

CEM-RL: Combining evolutionary and gradient-based methods for policy search *(Aloïs Pourchot et al)*

Learning and Planning with a Semantic Model *(Yi Wu et al)*

Deep learning

The Unreasonable Effectiveness of Deep Learning *(Richard Ngo)*

Large Scale GAN Training for High Fidelity Natural Image Synthesis *(Andrew Brock et al)*

Applications

Predicted Variables in Programming *(Victor Carbune et al)*

Alignment Newsletter #29

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

Highlights

Deep Imitative Models for Flexible Inference, Planning, and Control *(Nicholas Rhinehart et al)*: It's hard to apply deep RL techniques to autonomous driving, because we can't simply collect a large amount of experience with collisions in order to learn. However, imitation learning is also hard, because as soon as your car deviates from the expert trajectories that you are imitating, you are out of distribution, and you could make more mistakes, leading to accumulating errors until you crash. Instead, we can model the expert's behavior, so that we can tell when we are moving out of distribution, and take corrective action.

They split up the problem into three different stages. First, they generate a set of *waypoints* along the path to be followed, which are about 20m away from each other, by using A* search on a map. Next, they use model-based planning using an imitative model to generate a plan (sequence of states) that would take the car to the next waypoint. Finally, they use a simple PID controller to choose low-level actions that keep the car on target towards the next state in the plan.

The key technical contribution is with the imitative model, which is a probabilistic model P(s_{1:T}, G, φ), where φ is the current observation (eg. LIDAR), s_{1:T} is the planned trajectory, and G is a goal. We can learn P(s_{1:T} | φ) from expert demonstrations. The goal G can be anything for which you can write down a specification P(G | s_{1:T}, φ). For example, if you simply want to reach a waypoint, you can use the normal distribution on the distance between the final state s_T and the waypoint. You can also incorporate a hand-designed cost on each state.

They evaluate in simulation on a static world (so no pedestrians, for example). They show decent transfer from one map to a second map, and also that they can avoid artificially introduced potholes at test time (despite not seeing them at training time), simply by adding a cost on states over a pothole (which they can take into account because they are performing model-based planning).

Rohin's opinion: I really like this paper, it showcases the benefits of both model-based planning and imitation learning. Since the problem has been decomposed into a predictive model, a goal G, and a planner, we can edit G directly to get new behavior at test time without any retraining (as they demonstrate with the pothole experiment). At the same time, they can get away with not specifying a full reward function, as many features of good driving, like passenger comfort and staying in the correct lane, are learned simply by imitating an expert.

That said, they initially state that one of their goals is to learn from offline data, even though offline data typically has no examples of crashes, and "A model ignorant to the possibility of a crash cannot know how to prevent it". I think the idea is that you never get into a situation where you could get in a crash, because you never deviate from expert behavior since that would have low P(s_{1:T} | φ). This is better than model-based planning on offline data, which would consider actions that lead to a crash and have no idea what would happen, outputting garbage. However, it still seems that situations could arise where a crash is imminent, which don't arise much (if at all) in

the training data, and the car fails to swerve or brake hard, because it hasn't seen enough data.

Interpretability and Post-Rationalization *(Vincent Vanhoucke)*: Neuroscience suggests that most explanations that we humans give for a decision are post-hoc rationalizations, and don't reflect the messy underlying true reasons for the decision. It turns out that decision making, perception, and all the other tasks we're hoping to outsource to neural nets are inherently complex and difficult, and are not amenable to easy explanation. We can aim for "from-without" explanations, which post-hoc rationalize the decisions a neural net makes, but "from-within" explanations, which aim for a mechanistic understanding, are intractable. We could try to design models that are more interpretable (in the "from-within" sense), but this would lead to worse performance on the actual task, which would hurt everyone, including the people calling for more accountability.

Rohin's opinion: I take a pretty different view from this post -- I've highlighted it because I think this is an important disagreement that's relevant for alignment. In particular, it's not clear to me that "from-within" interpretability is doomed -- while I agree that humans basically only do "from-without" rationalizations, we also aren't able to inspect a human brain in the same way that we can inspect a neural net. For example, we can't see the output of each individual neuron, we can't tell what input would each neuron would respond maximally to, and we can't pose counterfactuals with slightly different inputs to see what changes. In fact, I think that "from-within" interpretability techniques, such as Building Blocks of Interpretability have already seen successes in identifying biases that image classifiers suffer from, that we wouldn't have known about otherwise.

We could also consider whether post-hoc rationalization is sufficient for alignment. Consider a thought experiment where a superintelligent AI is about to take a treacherous turn, but there is an explainer AI system that post-hoc rationalizes the output of the AI that could warn us in advance. If the explainer AI only gets access to the output of the superintelligent AI, I'm very worried -- it seems way too easy to come up with some arbitrary rationalization for an action that makes it seem good, you'd have to be have a much more powerful explainer AI to have a hope. On the other hand, if the explainer AI gets access to all of the weights and activations that led to the output, it seems more likely that this could work -- as an analogy, I think a teenager could tell if I was going to betray them, if they could constantly eavesdrop on my thoughts.

Technical AI alignment

Learning human intent

Deep Imitative Models for Flexible Inference, Planning, and Control *(Nicholas Rhinehart et al)*: Summarized in the highlights!

Addressing Sample Inefficiency and Reward Bias in Inverse Reinforcement Learning *(Ilya Kostrikov et al)*: Deep IRL algorithms typically work by training a discriminator that distinguishes between states and actions from the expert from states and actions from the learned policy, and extracting a reward function from the discriminator. In any environment where the episode can end after a variable number of timesteps, this assumes that the reward is zero after the episode ends. The reward function from the

discriminator often takes a form where it must always be positive, inducing a survival incentive, or a form where it must always be negative, inducing a living cost. For example, GAIL's reward is always positive, giving a survival incentive. As a result, *without any reward learning at all* GAIL does better on Hopper than behavioral cloning, and fails to learn on a reaching or pushing task (where you want to do the task as quickly as possible, so you want the living cost). To solve this, they learn an "absorbing state reward", which is a reward given after the episode ends -- this allows the algorithm to learn for itself whether it should have a survival incentive or living cost.

They also introduce a version that keeps a replay buffer of experience and uses an off-policy algorithm to learn from the replay buffer in order to improve sample efficiency.

Rohin's opinion: The key insight that rewards are *not* invariant to additions of a constant when you have variable-length episodes is useful and I'm glad that it's been pointed out, and a solution proposed. However, the experiments are really strange -- in one case (Figure 4, HalfCheetah) their algorithm outperforms the expert (which has access to the true reward), and in another (Figure 5, right) the blue line implies that using a uniformly zero reward lets you achieve around a third of expert performance (!!).

Interpretability

Interpretability and Post-Rationalization *(Vincent Vanhoucke)*: Summarized in the highlights!

Sanity Checks for Saliency Maps *(Julius Adebayo et al)*

Adversarial examples

Spatially Transformed Adversarial Examples *(Chaowei Xiao et al)* (summarized by Dan H): Many adversarial attacks perturb pixel values, but the attack in this paper perturbs the pixel locations instead. This is accomplished with a smooth image deformation which has subtle effects for large images. For MNIST images, however, the attack is more obvious and not necessarily content-preserving (see Figure 2 of the paper).

Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation *(Chaowei Xiao et al)* (summarized by Dan H): This paper considers adversarial attacks on segmentation systems. They find that segmentation systems behave inconsistently on adversarial images, and they use this inconsistency to detect adversarial inputs. Specifically, they take overlapping crops of the image and segment each crop. For overlapping crops of an adversarial image, they find that the segmentation are more inconsistent. They defend against one adaptive attack.

Uncertainty

On Calibration of Modern Neural Networks *(Chuan Guo et al.)* (summarized by Dan H): Models should not be unduly confident, especially when said confidence is used for decision making or downstream tasks. This work provides a simple method to make models more calibrated so that the confidence estimates are closer to the true correctness likelihood. (For example, if a calibrated model predicts “toucan” with 60% confidence, then 60% of the time the input was actually a toucan.) Before presenting

their method, they observe that batch normalization can make models less calibrated, while unusually large weight decay regularization can increase calibration. However, their proposed approach to increase calibration does not impact accuracy or require substantive model changes. They simply adjust the temperature of the softmax to make the model's "confidence" (here the maximum softmax probability) more calibrated. Specifically, after training they tune the softmax temperature to minimize the cross entropy (negative average log-likelihood) on validation data. They then measure model calibration with a measure which is related to the Brier score, but with absolute values rather than squares.

Dan H's opinion: Previous calibration work in machine learning conferences would often to focus on calibrating regression models, but this work has renewed interest in calibrating classifiers. For that reason I view this paper highly. That said, this paper's evaluation measure, the "Expected Calibration Error" is not a proper scoring rule, so optimizing this does not necessarily lead to calibration. In their approximation of the ECE, they use equally-wide bins when there is reason to use adaptively sized bins. Consequently I think Nguyen and O'Connor Sections 2 and 3 provide a better calibration explanation, better calibration measure, and better estimation procedure. They also suggest using a convex optimization library to find the softmax temperature, but at least libraries such as CVXPY require far more time and memory than a simple softmax temperature grid search. Finally, an understandable limitation of this work is that it assumes test-time inputs are in-distribution, but when inputs are out-of-distribution this method hardly improves calibration.

Miscellaneous (Alignment)

AI Alignment Podcast: On Becoming a Moral Realist with Peter Singer *(Peter Singer and Lucas Perry)*: There's a fair amount of complexity in this podcast, and I'm not an expert on moral philosophy, but here's an *oversimplified* summary anyway. First, in the same way that we can reach mathematical truths through reason, we can also arrive at moral truths through reason, which suggests that they are true facts about the universe (a moral realist view). Second, preference utilitarianism has the problem of figuring out which preferences you want to respect, which isn't a problem with hedonic utilitarianism. Before and after the interview, Lucas argues that moral philosophy is important for AI alignment. Any strategic research "smuggles" in some values, and many technical safety problems, such as preference aggregation, would benefit from a knowledge of moral philosophy. Most importantly, given our current lack of consensus on moral philosophy, we should be very wary of locking in our values when we build powerful AI.

Rohin's opinion: I'm not convinced that we should be thinking a lot more about moral philosophy. While I agree that locking in a set of values would likely be quite bad, I think this means that researchers should not hardcode a set of values, or create an AI that infers some values and then can never change them. It's not clear to me why studying more moral philosophy helps us with this goal. For the other points, it seems not too important to get preference aggregation or particular strategic approaches exactly perfect as long as we don't lock in values -- as an analogy, we typically don't argue that politicians should be experts on moral philosophy, even though they aggregate preferences and have large impacts on society.

Near-term concerns

Fairness and bias

A new course to teach people about fairness in machine learning *(Sanders Kleinfeld)*: Google has added a short section on fairness to their Machine Learning Crash Course (MLCC).

Privacy and security

Secure Deep Learning Engineering: A Software Quality Assurance Perspective *(Lei Ma et al)*

Other progress in AI

Reinforcement learning

Open sourcing TRFL: a library of reinforcement learning building blocks *(Matteo Hessel et al)* (summarized by Richard): DeepMind is open-sourcing a Tensorflow library of "key algorithmic components" used in their RL agents. They hope that this will allow less buggy RL code.

Richard's opinion: This continues the trend of being able to easily implement deep learning at higher and higher levels of abstraction. I'm looking forward to using it.

CURIOUS: Intrinsically Motivated Multi-Task, Multi-Goal Reinforcement Learning *(Cédric Colas et al)* (summarized by Richard): This paper presents an intrinsically-motivated algorithm (an extension of Universal Value Function Approximators) which learns to complete multiple tasks, each parameterised by multiple “goals” (e.g. the locations of targets). It prioritises replays of tasks which are neither too easy nor too hard, but instead allow maximal learning progress; this also help prevent catastrophic forgetting by refocusing on tasks which it begins to forget.

Richard's opinion: While I don’t think this paper is particularly novel, it usefully combines several ideas and provides easily-interpretable results.

Deep learning

Discriminator Rejection Sampling *(Samaneh Azadi et al)*: Under simplifying assumptions, GAN training should converge to the generator modelling the true data distribution while the discriminator always outputs 0.5. In practice, at the end of training the discriminator can still distinguish between images from the generator and images from the dataset. This suggests that we can improve the generated images by only choosing the ones that the discriminator thinks are from the dataset. However, if we use a threshold (rejecting all images where the discriminator is at least X% sure it comes from the generator), then we no longer model the true underlying distribution, since some low probability images could never be generated. They instead propose a rejection sampling algorithm that still recovers the data distribution under strict assumptions, and then relax those assumptions to get a practical algorithm, and show that it improves performance.

Meta learning

Meta-Learning: A Survey *(Joaquin Vanschoren)* (summarized by Richard): This taxonomy of meta-learning classifies approaches by the main type of meta-data they learn from:

1. Evaluations of other models on related tasks

2. Characterisations of the tasks at hand (and a similarity metric between them)

3. The structures and parameters of related models

Vanschoren explores a number of different approaches in each category.

Critiques (AI)

The 30-Year Cycle In The AI Debate *(Jean-Marie Chauvet)*

News

Introducing Stanford's Human-Centered AI Initiative *(Fei-Fei Li et al)*: Stanford will house the Human-centered AI Initiative (HAI), which will take a multidisciplinary approach to understand how to develop and deploy AI so that it is robustly beneficial to humanity.

Rohin's opinion: It's always hard to tell from these announcements what exactly the initiative will do, but it seems to be focused on making sure that AI does not make humans obsolete. Instead, AI should allow us to focus more on the creative, emotional work that we are better at. Given this, it's probably not going to focus on AI alignment, unlike the similarly named Center for Human-Compatible AI (CHAI) at Berkeley. My main question for the author would be what she would do if we could develop AI systems that could replace all human labor (including creative and emotional work). Should we not develop such AI systems? Is it never going to happen?

Read more: How to Make A.I. That’s Good for People

Why do we like stories?

This is a linkpost for https://steemit.com/philosophy/@yushih/why-do-we-like-stories-part-1

Some people like music. Others like visual art. Yet regardless of difference in age, gender and cultural backgrounds, everyone likes hearing a story. Whether it’s fantasized ‘Harry Potter’, romantic ‘Pride and Prejudice’, tragic 'Hamlet', or a beloved classic 'Journey to The West', man and women around the world are enchanted by their spell. Story is one of the most universal art forms. Our cultures are built on myths, and religions are passed on through legends.

A curious person would ask: Why do we like stories?

What Is a Story?

Before we try to find the origin of stories, we might first want to know what a story is. A cogent definition of story is given by Randy Olson, a scientist turned filmmaker who now teaches other scientists how to tell stories. He created a simple method for telling stories, called ABT—which stands for “And, But, Therefore”. These three words capture the basic structure of a story. For example, we can tell the story of "The Wizard of Oz" : "A little girl living on a farm in Kansas AND her life is boring, BUT one day a tornado sweeps her away to the land of Oz, THEREFORE she must undertake a journey to find her way home." [1]

This is idea is not entirely new: Aristotle, the philosopher and one of the first story analyzers, recognized that every story contains a three-act structure: Beginning, Middle, and End. The structure roughly corresponds to Olson’s "And", "But" and "Therefore". However, the advantages of Olson's idea are that the words are simple (they are among the most used words in English) and that each word has a meaning signifying its function in the story. "AND" connects relevant information to introduce the story; "BUT" brings in conflict; "Therefore" resolves the conflict and concludes the story.

Science and Story

While teaching storytelling to scientists over the years, Olson recognized that there is a similarity between ABT and a scientific paper. Most scientific papers have a structure of: Introduction, Method, Results, and Discussion. If we compare this to the structure of a story (ABT), we will find that "And" roughly corresponds to Introduction, "But" to Method, "Therefore" to Results and Discussion. This shows that storytelling is related to problem-solving.

Hypothesis

We can now come back to the question of why we like stories. My hypothesis is that storytelling began as a means of transferring problem-solution knowledge between people.

Think about a savage man who comes into contact with a tiger in the jungle. Running for his life, he spotted a cave and rushes inside it to avoid becoming the tiger's meal. Upon returning to the tribe, the savage starts relating the experience, including the location and time he met the tiger. This describes the introduction to the problem which is the "AND". After this, he describes what he did to avoid the danger: He escaped to a secure cave. This is the stage when he is trying to find a solution to the problem: the "BUT". Finally, he explains how he solved the problem by entering a cave to avoid danger: the “Therefore”.

Storytelling seems to be an innate trait of us. This points to an evolutionary explanation.

All Life is Problem Solving

When we see that a dog's nose resembles a human's nose, we can assume that they derive from a common ancestor. This resemblance, in evolutionary terms, is called 'homology'. The philosopher Karl Popper, in his book, proposed that we should regard the problem-solving ability as homologous across all species. That is: "All Life is Problem Solving". [2]

Life must adapt to its environment. The environment is not designed to ensure an organism's survival. Problems arises, and life must have solutions to them. Thus, to survive, the organism must have knowledge of problem-solutions. Most are not conscious knowledge, but are more like instinct stored inside of the biochemical constitution of the organism—in other words, the knowledge is in its body. For most of life, the knowledge is transferred through generations biologically via genetic inheritance and other processes. However, for humans, it could also be transferred by direct communication between people—through stories.

But Story is not True

Actually, what we consider a story today is mostly not about problem-solution knowledge, at least not factual knowledge. The original problem-solution knowledge of the ABT form, should be called a proto-story.

While thinking about the stories of today, I realized that most of them don't teach us about facts, but rather stimulate us emotionally. They are pleasurable to hear and have aesthetic values. Also, the most popular movies are often about good and evil (for example, Harry Potter and Lord of The Rings). This view is supported by the psychologist Jordan Peterson's lecture series that covers Bible stories, which he argues teach morality.

Finally, if we consider the three dimensions of universal value: truth, beauty, and goodness, we can see that stories represent both beauty and goodness. On the other hand, science/philosophy is a process that helps us arrive at some factual truth. However, because they are structurally similar, we can postulate that they both descend from proto-stories. Perhaps once they were viewed as the same, but then branched off on different paths.

Myth

In the relative small populations of ancient Greece, art and literature flourished. The Greeks also invented science and philosophy—the beginning of rationality. This may seem to be a coincidence, but I would argue that these two events are related.

Most primitive cultures have myths. In ancient Greek, storytelling matured into poetry. Aesthetically, it has reached a height that after two millennia, it is still included in the curricula and influenced today's culture. The Greek citizens attend plays as regular entertainment. They even have poetry competitions. It's not hard to imagine that the poets started to pay attention to the craft of the stories themselves, they have even tried to improve upon them.

I suspect that this attitude—critically examine the stories and try to improve upon them—is also the reason why they were able to invent science and philosophy. While poets try to improve their stories in the context of aesthetics and morality, at one point in history, some people started trying to truly understand the world. These people become philosophers or scientists. As Popper has wrote: "scientific theories originate from myths, and that a myth may contain important anticipations of scientific theories."

Language

The psychologist Karl Buhler distinguished three functions of language. The first two are found in both animal and humans. The third one is unique to humans.

1. Self-expressive function: the organism expresses its internal state
2. signaling function: animals exchange crude information between each other
3. descriptive function: since human started to become more conscious, they can use language to describe complex information, which represents facts in the world.

   Popper added another function:

4. the critical function: "the critical discussion of the truth or falsity of propositions. " In this way, humans can critically select and modify the descriptions, hence improve upon their knowledge.

I would argue there is another function: the (proto-)storytelling function, or the function that conveys problem-solution knowledge which has the form of the ABT structure. In my view, evolutionarily, the storytelling function came about in between the third and fourth function. Or to be more precise, it can incorporated to the third function. As we now know, problem-solving is an essential part of every species hence also humans. The descriptive function can be seen as an auxiliary of the problem-solving transferring function, since we describe things in order to solve problems. This is more consistent with the evolutionary theory of knowledge.

References

[1] Connection: Hollywood Storytelling Meets Critical Thinking - Randy Olson

[2] All Life is Problem Solving - Karl Popper

Maps of Meaning: Abridged and Translated

Jordan Peterson has come up in the rationalsphere before; SSC reviewed his recent book 12 Rules for Life, which caused me to read it; Jacobian wrote about The Jordan Peterson Mask, and Robin Hanson reviewed his major scholarly work, Maps of Meaning. A key line from Hanson's review:

> In sum, Peterson comes across as pompous, self-absorbed, and not very self-aware. But on the one key criteria by which such a book should most be judged, I have to give it to him: the book offers insight.

So this article is my attempt to distill the core insight I found in Maps of Meaning. One reason I titled this "abridged" is because Peterson gives excellent summaries of his sections, which I will often just reprint fully. "Translated" is because he goes about his case much differently than I would; understandable, given the difference between our audiences. Peterson spends much of the book establishing plausibility that many different cultures have similar myths, and explaining what they represent using his terminology, whereas I am not moved by archaic human universality; even if all ancient cultures believed that the Sun revolved around an unmoving Earth, I want to believe in modern astronomy. To the extent that his subject matter is human psychology, even if all ancient cultures had the same view of what humans were like, I want to focus on what WEIRD people are like. But, thankfully, Peterson is primarily making an argument for a better understanding of *progress*, not obedience to the past. First I'll try to explain the mythic perspective, and then Peterson's characterization of the human condition, and then some commentary.

WHAT A MYTH IS

> The world can be validly construed as forum for action, or as place of things.
>
> The former manner of interpretation--more primordial, and less clearly understood--finds its expression in the arts of humanities, in ritual, drama, literature, and mythology. The world as a forum for action is a place of value, a place where all things have meaning. This meaning, which is shaped as a consequence of social interaction, is implication for action, or--at a higher level of analysis--implication for the configuration of the interpretive schema that produces or guides action.
>
> The latter manner of interpretation--the world as place of things--finds its formal expression in the methods and theories of science. Science allows for increasingly precise determination of the consensually validatable properties of things, and for efficient utilization of precisely determined things as tools (once the direction such use is to take has been determined, through application of more fundamental narrative processes).
>
> No complete world-picture can be generated without use of both modes of construal. The fact that one mode is generally set at odds with the other means only that the nature of their respective domains remains insufficiently discriminated. Adherents of the mythological worldview tend to regard the

> statements of their creeds as indistinguishable from empirical “fact,” even though such statements were generally formulated long before the notion of objective reality emerged. Those who, by contrast, accept the scientific perspective--who assume that it is, or might become, complete--forget that an impassable gulf currently divides what is from what should be.

Some of the language here seems obscure on first reading--what exactly is meant by “forum for action,” or “consensually validatable”?--but eventually seemed sensible to me. By ‘forum for action,’ he means the agents that take actions have their worldview molded around the constraints of determining the right actions to take. To use one of Peterson’s later examples, consider a rat in a cage it has explored well. Add a foreign object to the cage, like a cube of iron; the rat will initially flee, then as no danger presents itself, it will cautiously approach and inspect the object, attempting to figure out what it is good for. Can it be eaten? Used as bedding? If this exploration reveals the object is useless, it will ignored, as it distracts from the bits of the environment that are relevant to deciding what actions to take.

But, of course, useless is a two-place word. The cube of iron may be useless_rat without being useless_vaniver; there might be useful actions I could do with that cube, like fiddle with it or toss it or decorate with it or throw it away or store it. This example so far has focused on possibilities in a way that made the values implicit--while the cube might be too large for the rat to swallow it, it’s small enough that I could, but I didn’t include that in my list of actions because none of my goals are advanced by eating raw iron.

So the mythological worldview is one where the motivational role of beliefs takes center stage. But do we really have to call this “myth”? People frequently talk about “narrative” in ways that capture this ‘forum for action’ or ‘motivational’ business. It seems like we could discard the word “myth”: calling it narrative might just be fine, and makes clearer the relevance of the myths that describe “where we are” and “how we got here” as opposed to just “how to behave.” For the rest of this post, I’ll use ‘myth’ and ‘narrative’ interchangeably, but default to ‘myth’ because it’s Peterson’s language.

There seem to be two important takeaways from this section of the book:

First, functional humans *require* a motivational worldview, and the type signature of a motivational worldview is different than the type signature of the outputs of science. That doesn’t mean scientific knowledge isn’t useful for building that motivational worldview, it just means that there’s a gap between the outputs of the scientific process and the inputs of your ‘tastes’ that has to be filled by something. Reading through Is Humanism A Religion Substitute with this lens, I come away with “yep, Humanism counts as a motivational worldview and can be understood through the lens of myth.” Reading through Raised in Technophilia, it’s easy to see how science (seeing the world as a place of things) is insufficient to reach that particular viewpoint, and how the mythological mode of thinking captures the cultural transmission that’s happening. Science just describes changes that are happening; it is technophilia that identifies those changes as *progress* and the people making them as the Good Guys.

Second, historical cultures operated *prior to* the formation of the materialist / reductionist / naturalist / empiricist paradigm, and so their claims will be misunderstood if parsed in the empiricist language instead of in the mythical language. The core job of cultures is to teach their members how to properly integrate into the society, and so their stories are about that, rather than about what actually

happened or how the world actually works. Stories and myths serve as catalogs of situations and examples of how to behave (or how not to behave) in those situations.

So suppose we agree that people need ‘purpose’ or ‘meaning’ or something similar, and that the type signature for this is ‘narrative’ or ‘myth’ or something similar. So what?

ORDER AND CHAOS

Peterson is mostly concerned with the myth of the hero, and spends most of the book discussing it. This is actually somewhat remarkable, given my presentation of the last chapter: one might assume that the book is like the catalog of human universals, and so collects myths that cover all of the important behaviors. That’s closer to what’s happening in Twelve Rules for Life; each of the rules is associated with stories that are important to the meaning and interpretation of that rule, and the combination of rules paints a picture of something like a whole human being. In Maps of Meaning, Peterson is instead trying to point to the core psychological need for myth, and the core myth that addresses this psychological need. Peterson spends much of the later bit of the book on the atrocities of the twentieth century; his desire to understand how they were possible (and, ideally, preventable) led to his discovery of the importance of myth, in part because this core myth ties into his diagnosis. The myth, in brief:

The world begins in a state of undifferentiated unity or unconsciousness; everything is the same because differences don’t exist yet. Then there is a separation, into the known and predictable (“order”) and the unknown and unpredictable (“chaos”). Order is secure and familiar; Chaos is dangerous and promising. The forces of order become insufficient to withstand the forces of chaos. An exploratory process (“the hero”) meets with chaos, defeats it, and transmutes it into order.

This is, in some sense, a presentation of the human condition--beginning in ignorance, learning some things, becoming self-aware enough to notice the difference between what is known and what is unknown, and deliberately learning about the unknown despite risks and discomfort and change. The myth typically includes a history of the gods that created the world, creating a mirror between the creation of an individual’s map and the creation of the external territory. Even beyond the individual process of learning and development, it describes the cultural process of learning and development; the society has some orderly way of dealing with what it knows, but the universe is larger than the society and so may change unpredictably (from the society’s viewpoint), causing decay and despair (another typical element of these myths) and the hero, by grappling with chaos, reforms the society and restores order and prosperity. Peterson often uses literary, figurative, and mythical language, in a way that I’m trying to avoid doing here, in a way that possibly makes his point easier to miss. For example, he connects ‘order’ with masculinity and ‘chaos’ with femininity, because this is so often done in historical myths, and one can see the connections. The unknown, containing many possibilities, is like the mother of many children. Tiamat (the feminine dragon of chaos) or her equivalent is depicted as birthing many different monsters. But while referring to the cluster of order-known-masculine and chaos-unknown-feminine lets you see the similarities across myths, it makes the reductionist project of separating out the distinct elements harder. If we load the core myth of the hero with all of these different features, then we can’t really point to a specific psychological need that this matches instead of ‘the human condition.’ It is not clear to me how much Peterson sees this holism as critical, as opposed to simply a

feature of accumulated traditions. When faced with the wisdom of millenia, it is somewhat arrogant to say “ah, this lesson that I inferred from this myth is the *final* lesson to be inferred from that myth.”

But to some extent, this is the problem of psychology and self-understanding. Once a piece of the puzzle moves from chaos to order, there is yet more chaos to incorporate. Peterson makes a big deal out of how the Egyptian heroic deity myth includes a detail that the Babylonian heroic deity myth doesn’t (and how the creation and inclusion of this detail was the product of heroic cultural activity). In the Babylonian tradition, the king is associated with Marduk, youngest of the gods, who defeats the world-threatening Tiamat and becomes king of the gods. In the Egyptian tradition, the pharaoh is associated with Horus, child of Osiris, who defeats the improper king Set and *rescues his father from the underworld*, and becomes king of the gods. Osiris, in this telling, is the cultural practices of the past, Set is the way in which those practices have become maladaptive, and Horus is the process of discernment and discovery that creates the cultural practices of the present in continuity with those of the past while adapting to changing circumstances.

Peterson elaborates on the details and incorporates more myths (including Christianity and alchemy), but this is enough to capture the basic worldview and how it seems fundamentally compatible with the rationalist worldview. But instead of discussing more myths, I’ll jump straight to his diagnosis of the 20th century.

THE HERO AND THE ADVERSARY

> One of these “hostile brothers” or “eternal sons of God” is the mythological hero. He faces the unknown with the presumption of its benevolence--with the (unprovable) attitude that confrontation with the unknown will bring renewal and redemption. He enters, voluntarily, into creative “union with the Great Mother,” builds of regenerates society, and brings peace to a warring world.
>
> The other “son of God” is the eternal adversary. This “spirit of unbridled rationality,” horrified by his limited apprehension of the conditions of existence, shrinks from contact with everything he does not understand. This shrinking weakens his personality, no longer nourished by the “water of life,” and makes him rigid and authoritarian, as he clings desperately to the familiar, “rational,” and stable. Every deceitful retreat increases his fear; every new “protective law” increases his frustration, boredom, and contempt for life. His weakness, in combination with his neurotic suffering, engenders resentment and hatred for existence itself.
>
> The personality of the adversary comes in two forms, so to speak--although these two forms are inseparably linked. The fascist sacrifices his soul, which would enable him to confront change on his own, to the group, which promises to protect him from everything unknown. The decadent, by contrast, refuses to join the social world, and clings rigidly to his own ideas--merely because he is too undisciplined to serve as an apprentice. The fascist wants to crush everything different, and then everything; the decadent immolates himself, and builds the fascist from his ashes. The bloody excesses of the twentieth century, manifest most evidently in the culture of the concentration camp, stand as testimony to the desires of the adversary and as monument to his power.

Here, “unbridled rationality” means something closer to the “high modernism” of Seeing Like a State. The core argument, as I understand it, is that totalitarianism grows from too much or too little confidence in knowledge. Both the overconfident and the underconfident have nothing left to learn; the first because they think they already know it (or it’s useless), and the second because they have abdicated their ability to learn. By nature, groups are associated with the known, rather than the process of the knower.

Peterson’s antidote is the heroic myth, as gradually constructed and refined by the slow accumulation of Western culture, which built up a myth of individual divinity / moral worth / moral judgment that involves a weird combination of humility and confidence--the sort of humility that allows one to actually adapt to reality, and the sort of confidence that allows one to stand up to the group or make mistakes while venturing into the unknown.

While a nice story (especially because it prescribes things I like), I find myself not convinced of its relevance or success. It seems like an important part of the story of Peterson’s life, and useful to include as context for why things are shaped particular ways (“ah, that’s why this aspect of the adversary is emphasized and that other bit minimized; this is the view of the adversary one gets by looking at the Nazis and Soviets.”). I also suspect that people should attempt to embody the heroic myth. But I suspect systemic problems have systemic solutions--the hero doesn’t just perfect themselves, they reorder society, and Peterson’s attempt to heroically reorder society in a way that causes more people to be heroes (because they have a crisper, more accessible view of what heroism means) seems possibly mismatched to the problem. It seems like the question of “what is the optimal number of scientists?”--yes, on the margin I would increase the amount of scientific thinking done by everyone, but how much I would pay to increase that for various people seems like the actual question.

It’s also not at all obvious that 20th century totalitarianism is incompatible with heroism as he sketches it out. Would a society in which more people set out to use their judgment to integrate their experience of their environment and the past be more tolerant and peaceful, or more revolutionary and violent?

CONCLUSION

Overall, the book was worth reading for me, and gave me a better sense of what ‘narrative’ is and what it’s useful for, but is not one I’d confidently recommend. The second quarter (“Mythological Representation: the Constituent Elements of Experience”) and last quarter (on alchemy) can probably be skipped, or heavily skimmed. Reading the summaries also seems like it captures much of the value of the book, and reading this summary perhaps captures much of the value of reading those summaries.

There seems to be a brand of scientific ennui which is somewhat common in the rationality community, where one focuses on predictions and models and ends up detached from the value in everyday life or pessimistic about one’s own prospects, which it seems like Peterson’s brand of mythological thinking is well-poised to counteract. As an example, the person who is convinced that the only cause that matters now is AI safety, but who is also convinced that they can’t be of any value to AI safety, seems at risk to see themselves as worthless and fall into depression. But this doesn’t help the project of AI safety, and it doesn’t help the individual in question. Contrast to ‘clean your room,’ which both makes the world more orderly (and the

individual feel more powerful) on the object level, and teaches lessons about scale on the meta level (in that one should tackle challenges that are appropriately sized, and by doing so level up to be able to take on larger and more complicated challenges).

Continuing on the topic of AI alignment, it seems to me like thinking about narrative is useful in at least two ways. First, it seems to point towards the sort of self-awareness that's necessary to encode ideas like corrigibility, and second, to the extent that alignment requires understanding humans, models of human psychology that more closely connect with what's actually important to humans seem more likely to capture what's important in life. For some reason, I feel more optimistic about an AI that has a good sense of what someone's desired heroic journey is than an AI that has a good sense of what someone's utility function is, in part because the first feels like it captures more meta-values and is less likely to be a snapshot of current opinions.

It seems to me like an important part of being a complete human is understanding the role narrative plays in one's thinking, and how to take command of it. Myths, narratives, one's taste and perception of the world as a place of fear and hope, danger and opportunity, and actions to be pursued or avoided, are all necessary parts of the human experience. The more advanced person does narrative to themselves, instead of having narrative done to them, embodying the role of the hero by incorporating their judgment into the process.

The Valley of Bad Theory

An interesting experiment: researchers set up a wheel on a ramp with adjustable weights. Participants in the experiment then adjust the weights to try and make the wheel roll down the ramp as quickly as possible. The participants go one after the other, each with a limited number of attempts, each passing their data and/or theories to the next person in line, with the goal of maximizing the speed at the end. As information accumulates "from generation to generation", wheel speeds improve.

This produced not just one but two interesting results.

First, after each participant's turn, the researchers asked the participant to predict how fast various configurations would roll. Even though wheel speed increased from person to person, as data accumulated, their ability to predict how different configurations behave did *not* increase. In other words, performance was improving, but understanding was not.

Second, participants in some groups were allowed to pass along both data and theories to their successors, while participants in other groups were only allowed to pass along data. Turns out, the data-only groups performed better. Why? The authors answer:

> … participants who inherited an inertia- or energy- related theory showed skewed understanding patterns. Inheriting an inertia-related theory increased their understanding of inertia, but decreased their understanding of energy…

And:

> … participants' understanding may also result from different exploration patterns. For instance, participants who received an inertia-related theory mainly produced balanced wheels (Fig. 3F), which could have prevented them from observing the effect of varying the position of the wheel's center of mass.

So, two lessons:

1. Iterative optimization does not result in understanding, even if the optimization is successful.
2. Passing along theories can actually make both understanding and performance worse.

So… we should iteratively optimize and forget about theorizing? Fox not hedgehog, and all that?

Well… not necessarily. We're talking here about a wheel, with weights on it, rolling down a ramp. Mathematically, this system just isn't all that complicated. Anyone with an undergrad-level understanding of mechanics can just crank the math, in all of its glory. Take no shortcuts, double-check any approximations, do it right. It'd be tedious, but certainly not intractable. And then… then you'd understand the system.

What benefit would all this theory yield? Well, you could predict how different configurations would perform. You could say for sure whether you had found the best solution, or whether better configurations were still out there. You could avoid getting

stuck in local optima. Y’know, all the usual benefits of actually understanding a system.

But clearly, the large majority of participants in the experiment did not crank all that math. They passed along ad-hoc, incomplete theories which didn’t account for all the important aspects of the system.

This suggests a valley of bad theory. People with no theory, who just iteratively optimize, can do all right - they may not really understand it, they may have to start from scratch if the system changes in some important way, but they can optimize reasonably well within the confines of the problem. On the other end, if you crank all the math, you can go straight to the optimal solution, and you can predict in advance how changes will affect the system.

But in between those extremes, there’s a whole lot of people who are really bad at physics and/or math and/or theorizing. Those people would be better off just abandoning their theories, and sticking with dumb iterative optimization.

Things I Learned From Working With A Marketing Advisor

Epistemic Status: Opinions stated without justification

I've been getting a bunch of advice and help at LRI from a marketing/strategy expert, and it's been an education. She's been great to work with — she kicks my ass, in a good way. Basically, she takes my writing, rips it apart, and helps me put it back together again, optimized to make the organization look better. Every kind of writing, from professional emails to website copy to grant proposals, gets a makeover. I'm thinking of the experience as something of an introduction to the conventions of business/promotional communication, which are very different from the kinds of writing norms I'm used to.

Here are some of the general patterns I've been learning about, stated in my own words (and maybe mangled a little in translation).

Discretization

"People hate reading," she tells me.

Seriously? You're going to rip up my nice, fluent, carefully-written essay explaining my rationale and replace it with a table?

Yes. Yes we are.

She's not wrong, though. I've had the experience of meeting with executives after sending them a *two-page document,* worrying that I should have written something more comprehensive, and finding they didn't even read the two-pager. I learn best through text, but clearly not everyone does. So promotional content needs to make allowances for the skimmers, the glancers, the reading-avoidant.

Hence: tables. Headers. Bolding key phrases. Bullet points. Pictures and graphs. Logos. And, of course, slide decks.

Layout matters. If you cross your eyes until the page turns blurry and *don't read anything*, how does it look? Is it a wall of text? If so, you need to break it up.

The principle of discretization is *things should be broken up into separate, distinctive, consistently labeled parts.*

What things? Everything.

Your website has parts. Your five-year plan has parts. Your value proposition has parts.

LRI doesn't have a "product", but in companies that sell a product, your product has parts called "features." Even when the "product" is sort of an abstract, general thing like "we produce written reports", in order to make them legible as products, you have to have a list of distinct parts that each report contains.

Once you have parts, you need to get obsessive about *matching* and *parallelism*. Each part needs to have one, and only one, name, and you have to use the same name

everywhere. If your organization has Five Core Values, you don't use near-synonyms to talk about them — you wouldn't interchangeably talk about "single focus" or "narrow mission", you'd pick one phrase, and use *that phrase* everywhere. Matchy-matchy.

You match your website navigation links to your page headers. You match your website to your grant proposals, your slide decks, your email phrasing, *everything*. You put your logo on every-fucking-thing. It feels repetitious to *you*, but it just looks appropriately consistent to an outside observer.

When I was a child, I was into American Girl dolls. My favorite thing was the parallelism. Each doll had five books, with matching titles and themes — "Changes for Felicity", "Changes for Samantha", etc. Each book came with its own outfit and accessories. The accessories were even parallel-but-unique — each doll had her own historically-accurate school lunch, her own toys, and so on. Even more than I liked actually playing with my doll, I liked reading through the catalog and noticing all the parallels. Ok, maybe I was a weird kid.

Anyhow, marketing is *full of that stuff*. Separating things into parallel-but-unique, hyper-matchy parts. Same principle as tables of correspondences.

I suspect that what you're doing is reifying your ideas into "existence." (In something like Heidegger's sense). You translate a general sort of concept ("I think we should test drugs to see which ones make animals live longer") into something with a bunch of proper nouns and internal structure, and I think the result is the overall impression that now your organization *exists*, as a…thing, or a place, or a personage. Like, the difference between an idea (e.g. the general concept of lifespan studies) and an agent (LRI). It activates the "animist" part of your brain, the same part that believes that Facebook is a place or Russia is an agent, the part that feels differently about proper nouns from improper nouns.

(Proper nouns, btw, are another big thing in themselves, because of social proof. Just naming people or institutions in connection with your work — whether they be advisors or partners or employees or customers or collaborators or whatever — is legitimizing. And proper nouns are, themselves, "discrete parts.")

All this discretization imparts a sense of legitimacy. After discretizing my writing, it feels much more like "LRI exists as a thing" rather than "Sarah is proposing an idea" or "Sarah is doing some projects." Yeah, that's a spooky and subjective distinction, but I think it's probably a very basic marketing phenomenon that permeates the world around us. (Maybe it has a name I don't know.) I feel slightly weird about it, but it's a *thing.*

Confidence + Pleasantries = Business Etiquette

One thing that came as a big surprise to me is how confident language you can get away with in a professional, non-academic context.

For example, not phrasing requests as questions. "I look forward to hearing back." My instinct would be to worry that this was overly forward or rude; you're essentially assuming the ask; but people don't seem to mind.

Or removing all uncertain language. All the may's, mights, and coulds. How can you do that without making overstated or misleading claims? Well, it's tricky, but you can generally finagle it with clever rephrasing.

I'm used to assuming that the way you show respect is through reticence and reluctance to ask for too much. Especially when communicating with someone higher status than you. To my surprise, *really* assertive wording seems to get better results with business types than my previous, more "humble" email style (which works great for professors.)

So, how do you keep from sounding like a jerk when you're essentially bragging and making big requests? A lot of pleasantries. A lot of framing phrases ("as we talked about in our last conversation", "circling back", "moving forward", etc). Wishing them a good weekend/holiday/etc, hoping they're doing well, etc.

I'd previously noticed in office contexts how vital it is to just *keep your mouth making words smoothly* even when there's not a lot of information density to what you're saying.

Business "jargon" and "buzzwords" are unfairly maligned by people who aren't used to corporate culture. First of all, a lot of them originally referred to specific important concepts, and then got overused as generic applause lights — e.g. "disruptive innovation" is actually a *really useful idea* in its original meaning. But, second of all, it's honestly just handy to have stock phrases if you need to keep talking fluently without awkward pauses. People respond really well to fluency. Palantir's first exercise for all new employees is to give a software demo, which taught me that it is *really hard* to speak in public for five minutes without pausing to think of what to say next. Stock phrases help you reach for something to say without appearing hesitant or afraid.

I was trained on writing style guides from literary or journalistic contexts, like Strunk & White, which teach you to be relentless in removing cliches and using simple short Anglo-Saxon words wherever possible. Business language constantly violates those rules: it's full of cliches and unnecessary Latinate locutions. But I suspect there may actually be a function to that, in making you sound smoother, or setting the scene with comfortable and familiar wording before introducing new ideas. "Good writing" is original and vivid; a good (i.e. effective) business email may not be.

Where is my Flying Car?

This is a linkpost for
http://www.bayesianinvestor.com/blog/index.php/2018/10/14/where-is-my-flying-car/

Book review: Where Is My Flying Car? A Memoir of Future Past, by J. Storrs Hall (aka Josh).

If you only read the first 3 chapters, you might imagine that this is the history of just one industry (or the mysterious lack of an industry).

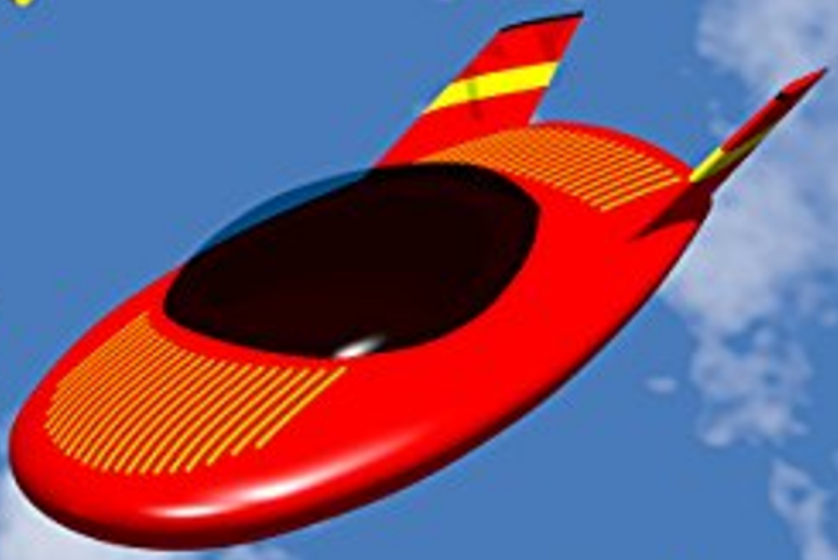
But this book attributes the absence of that industry to a broad set of problems that are keeping us poor. He looks at the post-1970 slowdown in innovation that Cowen describes in The Great Stagnation[1]. The two books agree on many symptoms, but describe the causes differently: where Cowen says we ate the low hanging fruit, Josh says it's due to someone "spraying paraquat on the low-hanging fruit".

The book is full of mostly good insights. It significantly changed my opinion of the Great Stagnation.

The book jumps back and forth between polemics about the Great Strangulation (with a bit too much outrage porn), and nerdy descriptions of engineering and piloting problems. I found those large shifts in tone to be somewhat disorienting - it's like the author can't decide whether he's an autistic youth who is eagerly describing his latest obsession, or an angry old man complaining about how the world is going to hell (I've met the author at Foresight conferences, and got similar but milder impressions there).

Josh's main explanation for the Great Strangulation is the rise of Green fundamentalism[2], but he also describes other cultural / political factors that seem related. But before looking at those, I'll look in some depth at three industries that exemplify the Great Strangulation.

Where Is My Flying Car?



A Memoir of Future Past

by J Storrs Hall

The good old days of Science Fiction

The leading SF writers of the mid 20th century made predictions for today that looked somewhat close to what we got in many areas, with a big set of exceptions in the areas around transportation and space exploration.

The absence of flying cars is used as an argument against futurists' ability to predict technology. This can't be dismissed as just a minor error of some obscure forecasters. It was a widespread vision of leading technologists.

Josh provides a decent argument that we should treat that absence as a clue to why U.S. economic growth slowed in the 1970s, and why growth is still disappointing.

Were those SF writers clueless optimists, making mostly random forecasting errors? No! Josh shows that for the least energy intensive technologies, their optimism was about right, and the more energy intensive the technology was, the more reality let them down.

Is it just a coincidence that people started worshiping energy conservation around the start of the Great Stagnation? Josh says no, we developed ergophobia - no, not the standard meaning of ergophobia: Josh has redefined it to mean fear of using energy.

Did flying cars prove to be technically harder than expected?

The simple answer is: mostly no. The people who predicted flying cars knew a fair amount about the difficulty, and we may have forgotten more than we've learned since then.

Josh describes, in more detail than I wanted, a wide variety of plausible approaches to building flying cars. None of them clearly qualify as low-hanging fruit, but they also don't look farther from our grasp than did flying machines in 1900.

How serious were the technical obstacles?

Air traffic control

Before reading this book, I assumed that there were serious technical problems here. In hindsight, that looks dumb.

Josh calculates that there's room for a million non-pressurized aircraft at one time, under current rules about distance between planes (assuming they're spread out evenly; it doesn't say all Tesla employees can land near their office at 9am). And he points out that seagull tornadoes (see this video) provide hints that current rules are many orders of magnitude away from any hard limits.

Regulators' fear of problems looks like an obstacle, but it's unclear whether anyone put much thought into solving them, and it doesn't look like the industry got far enough for this issue to be very important.

Skill

It seems unlikely that anywhere near as many people would learn to fly competently as have learned to drive. So this looks like a large obstacle for the average family, given 20th century technology.

But we didn't get close the point where that was a large obstacle to further adoption. And 21st century technology is making progress toward convenient ways of connecting competent pilots with people who want to fly, except where it's actively discouraged.

Cost

If the economic growth of 1945-1970 had continued, we'd be approaching wealth levels where people on a UBI ... oops, I mean on a national basic income could hope to afford an occasional ride in a flying Uber that comes to their door. At least if there were no political problems that drove up costs.

Weather

Weather will make flying cars a less predictable means than ground cars to get to a given destination. That seems to explain a modest fraction of people's reluctance to buy flying cars, but that explains at most a modest part of the puzzle.

Safety

> The leading cause of death among active pilots is ... motorcycle accidents.

I wasn't able to verify that, and other sources say that general aviation is roughly as dangerous as motorcycles. Motorcycles are dangerous enough that they'd likely be illegal if they hadn't been around before the Great Strangulation, so whether either of those are considered safe enough seems to depend on accidents of history.

People have irrational fears of risk, but there has also been a rational trend of people demanding more safety because we can now afford more safety. I expect this is a moderate part of why early SF writers overestimated demand for flying cars.

The liability crisis seems to have hit general aviation harder than it hit most other industries. I'm still unclear why.

> One of the more ironic regulatory pathologies that has shaped the world of general aviation is that most of the planes we fly are either 40 years old or homemade - and that we were forced into that position in the name of safety.

If the small aircraft industry hadn't mostly shut down, it's likely that new planes would have more safety features (airbags? whole-airplane parachutes?).

The flying car industry hit a number of speedbumps, such as WWII diverting talent and resources to other types of aviation, then a key entrepreneur being distracted by a patent dispute, and then was largely shut down by liability lawsuits. It seems like progress should have been a bit faster around 1950-1970 - I'm confused as to whether the industry did well then.

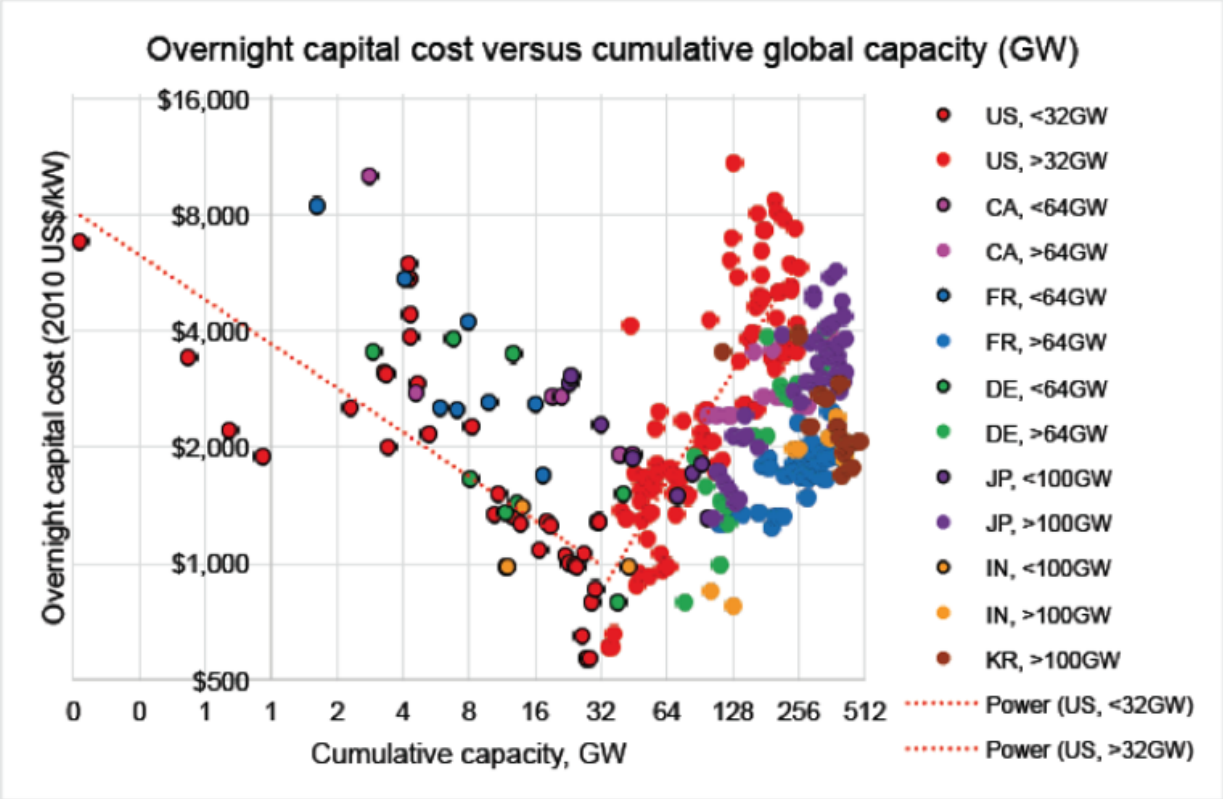
At any rate, it looks like liability lawsuits were the industry's biggest problem, and they combined with a more hostile culture and expensive energy to stop progress around 1980.

The book shifted my opinion from "those SF writers were confused" to "flying cars should be roughly as widespread as motorcycles". We should be close to having autopilots which eliminate the need for human pilots (and the same for motorcycles?), and then I'd consider it somewhat reasonable for the average family to have a flying car.

Nuclear Power

Josh emphasizes the importance of cheap energy for things such as flying cars, space travel, eradicating poverty, etc., and identifies nuclear power as the main technology that should have made energy increasingly affordable. So it seems important to check his claims about what went wrong with nuclear power.

He cites a study by Peter Lang, with this strange learning curve:



It shows a trend of costs declining with experience, just like a normal industry where there's some competition and where consumers seem to care about price. Then that trend was replaced by a clear example of cost disease[3]. I've previously blogged about the value of learning curves (aka experience curve effects) in forecasting.

This is pretty inconsistent with running out of low-hanging fruit, and is consistent with a broad class of political problems, including the hypothesis of hostile regulation, and also the hypothesis that nuclear markets were once competitive, then switched to having a good deal of monopoly power.

This is a pretty strong case that something avoidable went wrong, but leaves a good deal of uncertainty about what went wrong, and Josh seemed a little too quick to jump to the obvious conclusion here, so I investigated further[4]. I couldn't find anyone arguing that nuclear power hit technical problems around 1970, but then it's hard to find many people who try to explain nuclear cost trends at all.

This book chapter suggests there was a shift from engineering decisions being mostly made by the companies that were doing the construction, to mostly being determined by regulators. Since regulators have little incentive to care about cost, the effect seems fairly similar to the industry becoming a monopoly. Cost disease seems fairly normal for monopolies.

That chapter also points out the effects of regulatory delays on costs: "The increase in total construction time ... from 7 years in 1971 to 12 years in 1980 roughly doubled the final cost of plants."[5]

In sum, something went wrong with nuclear power. The problems look more political than technical. The resulting high cost of energy slowed economic progress by making some new technologies too expensive, and by diverting talent to energy conservation. And by protecting the fossil fuel industries, it caused millions of deaths, and maybe 174 Gt of unnecessary CO2 emissions (about 31% of all man-made CO2 emissions).

This book convinced me that I'd underestimated how important nuclear power could have been.

Nanotech

> So the technology of the Second Atomic Age will be a confluence of two strongly synergistic atomic technologies: nanotech and nuclear.

The book has a chapter on the feasibility of Feynman / Drexler style nanotech, which attempts to find a compromise between Drexler's excruciatingly technical Nanosystems and his science-fiction style Engines of Creation. That compromise will convince a few people who weren't convinced by Drexler, but most people will either find it insufficiently technical, or else hard to follow because it requires a good deal of technical knowledge.

Josh explains some key parts of why the government didn't fund research into the Feynman / Drexler vision of nanotech: centralization and bureaucratization of research funding, plus the Machiavelli Effect

- the old order opposes change, and beneficiaries of change "do not readily believe in new things until they have had a long experience of them."

Josh describes the mainstream reaction to nanotech fairly well, but that's not the whole story.

Why didn't the military fund nanotech? Nanotech would likely exist today if we had credible fears of Al Qaeda researching it in 2001. But my fear of a nanotech arms race exceeds my desire to use nanotech.

Many VCs would get confused by top academics who dismissed (straw-man versions of) Drexler's vision. But there are a few VCs such as Steve Jurvetson who understand Drexler's ideas well enough to not be confused by that smoke. With those VCs, the explanation is no entrepreneurs tried a sufficiently incremental path

Most approaches to nanotech require a long enough series of development steps to achieve a marketable product that VCs won't fund them. That's not a foolish mistake on VCs part - they have sensible reasons to think that some other company will get most of the rewards (how much did Xerox get from PARC's UI innovations?). Josh promotes an approach to nanotech that seems more likely to produce intermediate products which will sell. As far as I know, no entrepreneurs attempted to follow that path (maybe because it looked too long and slow?).

The patent system has been marketed as a solution to this kind of problem, but it seems designed for a hedgehog-like model of innovation, when what we ought to be incentivizing is a more fox-like innovation process.

Mostly there isn't a good system of funding technologies that take more than 5 years to generate products.

If government funding got this right during the golden age of SF, the hard questions should be focused more on what went right then, than on what is wrong with funding now. But I'm guessing there was no golden age in which basic R&D got appropriate funding, except when we were lucky enough for popular opinion to support the technologies in question.

Problems with these three industries aren't enough to explain the stagnation, but Josh convinced me that the problems which affected these industries are more pervasive, affecting pretty much all energy-intensive technologies.

Culture and politics

> Of all the great improvements in know-how expected by the classic science-fiction writers, competent government was the one we got the least.

I'll focus now on the underlying causes of stagnation.

Green fundamentalism and ergophobia are arguably sufficient to explain the hostility to nuclear power and aviation, but it's less clear how they explain the liability crisis or the stagnation in nanotech.

Josh also mentions a variety of other cultural currents, each of which explain some of the problems. I expect these are strongly overlapping effects, but I won't be surprised if they sound as disjointed as they did in the book.

It matters whether we fear an all-seeing god. From the book Big Gods: How Religion Transformed Cooperation and Conflict:

> In a civilization where a belief in a Big God is effectively universal, there is a major advantage in the kind of things you can do collectively. In today's America, you can't be trusted to ride on an airliner with a nail file. How could you be trusted driving your own 1000-horsepower flying car? ... The green religion, on the other hand, instead of enhancing people's innate conscience, tends to degrade it, in a phenomenon called "licensing." People who virtue-signal by buying organic products are more likely to cheat and steal

[6]

From Peter Turchin: when an empire becomes big enough to stop worrying about external threats to its existence, the cooperative "we're all in the same boat" spirit is

replaced by a "winner take all" mentality.

> the evolutionary pressures to what we consider moral behavior arise *only in non-zero-sum interactions*. In a dynamic, growing society, people can interact cooperatively and both come out ahead. In a static no-growth society, pressures toward morality and cooperation vanish;

Self deception is less valuable on a frontier where you're struggling with nature than it is when most struggles involve social interaction, where self-deception makes virtue signaling easier.

"If your neighbor is Saving the Planet, it seems somehow less valuable merely to keep clean water running".

"Technologies that provoke antipathy and promote discord, such as social networks, are the order of the day; technologies that empower everyone but require a background of mutual trust and cooperation, such as flying cars, are considered amusing anachronisms."

Those were Josh's points. I'll add these thoughts:

It's likely that cultural changes led competent engineers to lose interest in working for regulatory agencies. I don't think Josh said that explicitly, but it seems to follow fairly naturally from what he does say.

Josh refers to Robin Hanson a fair amount, but doesn't mention Robin's suggestion that increasing wealth lets us return to forager values. "Big god" values are clearly farmer values.

Mancur Olson's The Rise and Decline of Nations (listed in the bibliography, without explanation), predicted in 1982 that special interests would be an increasing drag on growth in stable nations. His reasoning differs a fair amount from Josh's, but their conclusions sound fairly similar.

Josh often focuses on Greens as if they're a large part of the problem, but I'm inclined to focus more on the erosion of trust and cooperation, and treat the Greens more as a symptom.

The most destructive aspects of Green fundamentalism can be explained by special interests, such as coal companies and demagogues, who manipulate long-standing prejudices for new purposes. How much of Great Strangulation was due to special interests such as coal companies? I don't know, but it looks like the coal industry would have died by 2000 (according to Peter Lang) if the pre-1970 trends in nuclear power had continued.

Green religious ideas explain hostility to energy-intensive technologies, but I have doubts about whether that would be translated into effective action. Greens could have caused cultural changes that shifted the best and the brightest away from wealth creation and toward litigation.

That attempt to attribute the stagnation mainly to Greens seems a bit weaker than the special interests explanation. But I remain very uncertain about whether there's a single cause, or whether it took several independent errors to cause the stagnation.

What now? I don't see how we could just turn on a belief in a big god. The book says we'll likely prosper in spite of the problems discussed here, but leaves me a bit gloomy

about achieving our full potential.

The book could use a better way of labeling environmentalists who aren't Green fundamentalists. Josh clearly understands that there are big differences between Green fundamentalists and people with pragmatic motives for reducing pollution or preserving parks. Even when people adopt Green values mostly for signaling purposes, there are important differences between safe rituals, such as recycling, and signals that protect the coal industry.

Yet standard political terminology makes it sound like attacks on the Greens signal hostility to all of those groups. I wish Josh took more care to signal a narrower focus of hostility.

Ironically for a book that complains about virtue signaling, a fair amount of the book looks like virtue signaling. Maybe that gave him a license to ignore mundane things like publicizing the book (I couldn't find a mention of the book on his flying car blog until 3 months after it was published).

Has the act of writing this review licensed me to forget about being effective? I'm a bit worried.

Miscellaneous comments and complaints

> It isn't perhaps realized just how much the war on cars contributed to the great stagnation - or how much flying cars could have helped prolong the boom.

Josh provides a good analysis of the benefits of near-universal car ownership, and why something similar should apply to flying cars. But he misses what I'll guess was the biggest benefit of cars - people applied for jobs for which they couldn't have previously managed to get to an interview. Company towns were significant in the 19th century - with downsides that bore some similarity to slavery, due to large obstacles to finding a job in another town. Better transportation and communications changed that.

He says "a century of climate change *in the worst case* might cost us as much as liability lawyers do now." He gets his estimate of the worst case from this GAO report. That's misleading about how we should evaluate the actual worst case. I'm not too clear how they got those numbers, but they likely mean something more like that there's a 95% chance that *according to some model*, climate change will do no more damage than lawyers. That still leaves plenty of room for the worst 1% of possible outcomes to be much worse than lawyers according to the model, and there's enough uncertainty in climate science that we should expect more than a 5% chance of the model erring on the optimistic side. Note also that it's not hard to find a somewhat respectable source that says climate change might cost over 20% of global GDP. I see other problems with his climate change comments, but they seem less important than his dismissal of the tail risks.

Josh reports that flying a plane causes him to think in far mode, much like our somewhat biased view of the future.

It's been a long time since I've flown a plane, but I don't recall that effect being significant. I find that a better way to achieve that experience is to hike up a mountain whose summit is above the clouds. Although there are relatively few places that have an appropriate mountain nearby, and it takes somewhat special timing where I live to do that.

While researching this review, I found this weird litigation story: Disney Sued for Not Building Flying “Star Wars” Car.

I often tend to side with technological determinist views of history, but this book provides some evidence against that. Just compare Uber with "Uber for planes" - it looks like there's a good deal of luck involved in what progress gets allowed.

Josh illustrates the Machiavelli Effect by an example of expert advice that fat is unhealthy, and he complains that the experts ignore Gary Taubes carbophobic counter-movement. Yet what I see is people on both sides of that debate focusing on interventions that are mostly irrelevant.

Josh points out that we can test the advice, and reports that he lost a good deal of weight after switching to a high-fat diet. Well, I tried a similar switch in 2012 from a low-fat diet to a high-fat diet, and it had no effect on my weight (and a terrible effect on my homocysteine and sdLDL, due to high saturated fat). The dietary changes the had the best effects on my weight were alternate day calorie restriction, cutting out junk food (mainly via paleo heuristics), and eating less kelp (which was depressing my thyroid via excess iodine).

He cites Scott Alexander in other contexts, but apparently missed this post pointing out serious flaws in Taubes' claims. Note also that Taubes reacted poorly to evidence against his theory.

Miscellaneous questions prompted by the book

The book hints that cultural beliefs have important influences on where smart people apply their talents. This mostly seems hard to analyze. Would Elon Musk be swayed by ergophobia or Green fundamentalism? That seems like the main example I can generate about a competent tech leader whose plans seem somewhat influenced by what popular beliefs about where technology should head. Tesla and SolarCity arguably fit a pattern of Musk being influenced by Green visions. But SpaceX looks more like pandering to the visions of ergophiles.

The book left me wondering: where does high modernism fit into this story? I see many similarities between high modernism and this book's notion of who the bad guys are. Yet high modernism started to crumble a bit before the worst parts of the Great Strangulation started (i.e. around 1970). The book hints at a semi-satisfying answer: Christianity and high modernism produced a decent balance of power where each ideology checked the others' excesses, but Green fundamentalism eroded the good aspects of high modernism while strengthening the worst aspects.

Did oil prices rise in the 1970s due to evidence that nuclear prices were rising? I can almost imagine OPEC being prescient enough to see that nuclear regulation saved them from important competition. The timing of OPEC's initial effects on the market seems to closely coincide with the nuclear industry developing cost disease. But I don't quite expect that OPEC leaders were that smart.

Another odd hypothesis: increasing mobility enabled people to move too easily to better jurisdictions. This scared lots of special interests (e.g. local governments, companies with a local monopoly, etc., whose power depended on captive customers), who reacted by advocating policies which reduced mobility (e.g. stifling transportation, encouraging home ownership instead of renting).

Quotes

I've only tried to summarize and analyze the more modest and basic parts of the book here. Some parts of the book are too strange for me to want to review. I will close with some quotes from them:

Hmmm. This might explain some of the book's peculiarities: "ideation recapitulates inebriation!".

"The human of the future will have more and better senses, be stronger and be adaptable to a much wider range of environments, and last but not least have the biosphere atom-rearranging capability built in. The human of the future need not have any ecological footprint at all."

His favorite form of renewable energy is nuclear: "In other words, if we start taking uranium out of seawater and use it for the entire world's energy economy, indeed a robustly growing energy economy, the concentration in seawater will not decline for literally millions of years."

"In the Second Atomic Age, Litvenenko would have gotten a text from his left kidney telling him that it had collected 26.5 micrograms of Polonium-210, and what would he like to do with it?"

He asks us not to call this a greenhouse: "The LEDs emit only the frequencies used by chlorophyll, so they are an apparently whimsical purple. The air is moist, warm, and has a significantly higher fraction of CO2 than natural air ... the plants do not need pesticides because insects simply can't get to them. ... you get something like 300 times as much lettuce per square foot of ground than the pre-industrial mule-and-plow dirt farmer. All you need is power, to have fresh local strawberries in January in the Yukon or in August in Antarctica."

And he likes tall buildings. I don't want to classify this comment:

> A ten-mile tower might have a footprint of a square mile and could house 40 million people. Eight such buildings would house the entire current population of the United States, leaving 2,954,833 square miles of land available for organic lavender farms.
>
> Compared to the skyhook (geostationary orbital tower), which is just barely possible even with the theoretical best material properties, a tower 100 km high is easy. Flawless diamond, with a compressive strength of 50 GPa, does not even need a taper at all for a 100 km tower; a 100-km column of diamond weights 3.5 billion newtons per square meter but can support 50 billion. Even commercially available polycrystalline synthetic diamond with advertised strengths of 5 GPa would work.
>
> A Weather Machine could probably double global GDP simply by regional climate control. ... You could make land in lots of places, such as Northern Canada and Russia, as valuable as California.

Um, don't forget the military implications which might offset that.

I used to be sort of comfortable with Reynolds numbers and lift-to-drag ratios, but this claim seems to be beyond my pay grade:

> Given the ridiculous wingspan and the virtually infinite Reynolds number, we might get a lift-to-drag ratio of 100; we would need 1 billion pounds of thrust.

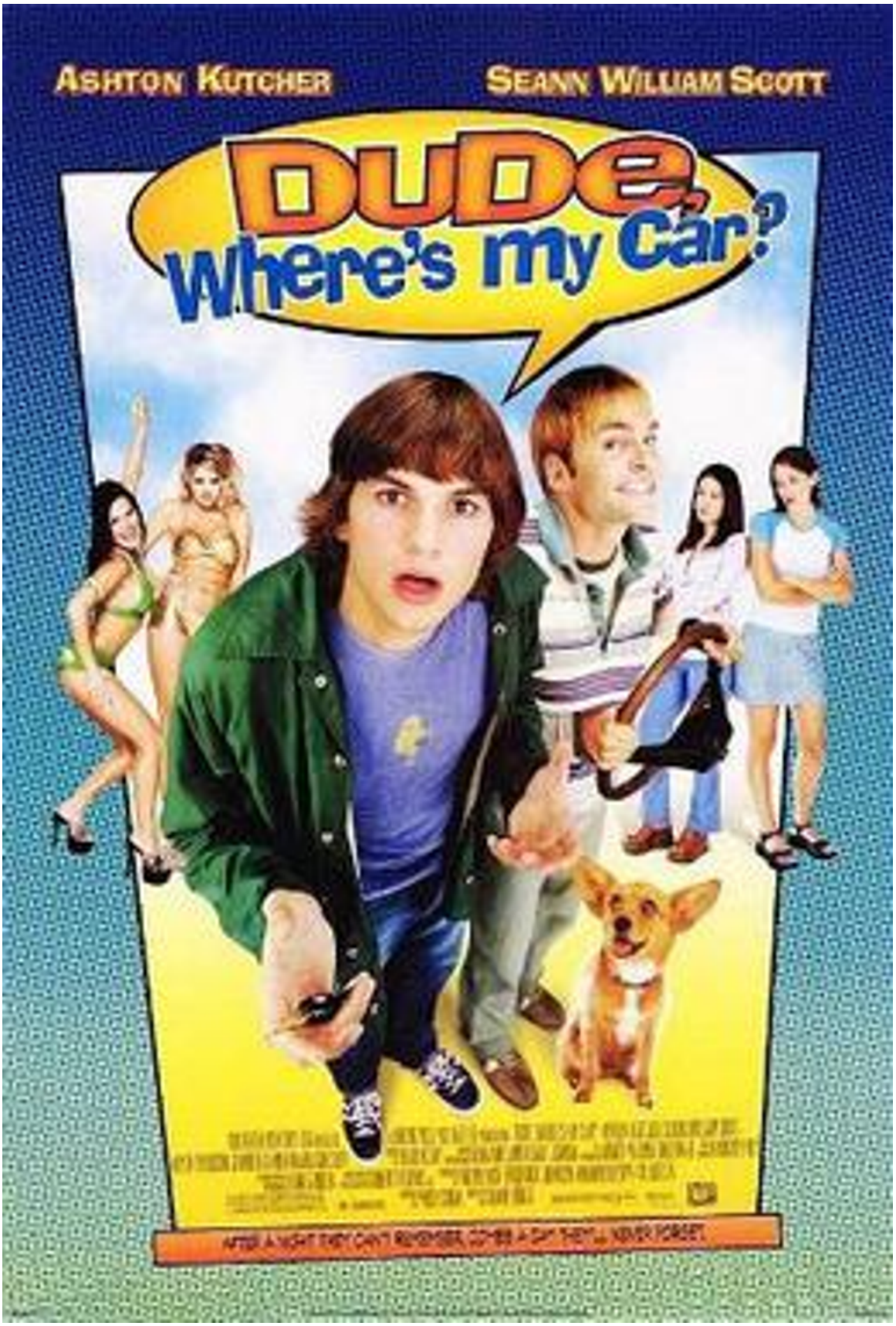
He's interested in cold fusion, but admits it's hard:

> But we would like a theory in which whenever some mechanism causes miracle 1 to happen, it *almost always* causes miracles 2 and 3. ... It seems at first blush that saying there might be a quantum coupling between phonons and some nuclear degree of freedom is indistinguishable from magic. But if you look closely, it's not completely insane.

I'll take his word on that for now, since "look closely" appears to require way more physics than I'm up for.

Biotech gets approximately one paragraph, including: "Expect Astro the talking dog before 2062. Expect to live long enough to see him."

> One of the hardest jobs that humans do, some well and some poorly, is management of other humans. One of the major reasons this is hard is that humans are selfish, unreasonable, fractious, and just plain ornery. ... On the other hand, managing robots with human-level competence will be falling-down easy. In the next couple of decades, robots will be climbing up the levels of competence to compete with humans at one job and another. Until they become spectacularly better, though, I suspect that the major effect will be to make management easier - perhaps so easy a robot could do it! Once we build trustworthy IQ 200 machines, only an idiot will trust any human to make any decision that mattered ...
>
> What then are we humans supposed to do?
>
> Don't look at me! We already know that only a fool would ask a human such an important question. Ask the machines.
>
> when someone invents a method of turning a Nicaragua into a Norway, extracting only a 1% profit from the improvement, they will become rich beyond the dreams of avarice and the world will become a much better, happier, place. Wise incorruptible robots may have something to do with it.


ASHTON KUTCHER
SEANN WILLIAM SCOTT
DUDE, Where's my Car?

Footnotes

[1] - I haven't read The Great Stagnation, so I'm commenting based on simple summaries of it. Based on what I know of Cowen, the books are of superficially similar quality. Cowen does an unusual amount of broad but shallow research, whereas Josh is less predictable about his research quality, but his research is often much deeper than Cowen's. E.g. for this book, it included learning how to fly, and buying a plane. That research alone likely cost him more money than he'll make from the book (not to mention hundreds of hours of his time), and it's not the only way in which his research is surprisingly deep.

[2] - not in quite the same sense as what people who call themselves Green fundamentalists mean, but pretty close. Both sides seems to agree that a key issue is whether industrial growth is good or bad.

Some of what Josh dislikes about the worst Greens:

> My own doubts came when DDT was introduced. In Guyana, within two years, it had almost eliminated malaria. So my chief quarrel with DDT, in hindsight, is that it has greatly added to the population problem.
>
> - Alexander King

[3] - at least in many countries. South Korea's nuclear costs have continued to decline. The variation in when cost disease hits suggests something other than engineering problems.

I got concerned about the lack of data from China. I couldn't find comparable Chinese data, so I used financial data from CGN Power Company (2011 data here, first half 2018 data here) to show, if my math is right, that CGN sold power at RMB0.3695 ($0.0558) / KWh in 2011 versus RMB0.2966 ($0.0448) / KWh in 2018, a decline of nearly 20%. I.e. no cost disease there.

Note: I own stock in CGN Power.

Josh claims that the Navy's nuclear power program avoided strangulation. Where can I get data about the cost trends there?

[4] - I've looked for anti-nuke arguments about the cost of nuclear power, and most seem to assume that cost disease is inevitable. A few look for signs that nuclear power has been treated unfairly, and focus on things like subsidies or carbon taxes.

It seems quite plausible that they start with the assumption that most wealth is a gift from Mother Nature, and conclude that most important conflicts are zero-sum struggles over who gets those gifts. They don't see anything that looks like taking resources away from nuclear power, and conclude that nuclear power has been regulated fairly.

Let me suggest an analogy: imagine the early days of the dot-com boom, when the benefits of Google search were not widely understood. Imagine also a coalition of music distributors, and people who are devoted to community-building via promoting social interaction in local libraries. Such a coalition might see Google as a threat, and point to the risks that Google would make porn more abundant. Such a coalition might well promote laws requiring Google to check each search result for porn (e.g. via manual inspection, or by only indexing pages of companies who take responsibility for keeping porn off their sites). It would be obvious that Google needs to charge a

moderately high subscription fee for its search - surely the new rules would only increase the subscription fees by a small fraction. [It actually seemed obvious to most hypertext enthusiasts up through about 1995 that a company like Google would need to charge users for its service.] Oh, and Xanadu has some interesting ideas for how to use micropayments to more easily charge for that kind of service - maybe Google can run under Xanadu?

A person who had no personal experience of benefiting from Google might not notice much harm from such a regulation, or might assume it has a negligible effect on Google's costs. And someone who imagines that Mother Nature is the primary source of free lunches is likely to seriously underestimate the benefits of Google.

I've seen occasional hints that people attribute the cost increases to valuable safety measures that had been missing from early reactors, but I haven't found anyone saying that who seems aware of the risks of keeping other energy sources in business. So I'm inclined to treat that the way I treat concerns about the safety of consumers pumping gas, or the dangers of caffeinated driving.

[5] - note that the high inflation of the time complicates that picture. A more simplified model would go like this: imagine 0% inflation, and the company borrows money at an interest rate of 5%. Then a 5-year delay causes the cost of capital to rise 27.6% (1.05^5). Cost of capital is one of the larger costs of nuclear power, so the delays alone look sufficient to turn nuclear power from quite competitive to fairly uncompetitive.

I expect that people who are unfamiliar with finance will underestimate the significance of this.

[6] - Josh says this is an example of how science works pretty well: social scientists are likely quite biased against this conclusion, but keep upholding it.

A compendium of conundrums

This is a linkpost for http://thinkingcomplete.blogspot.com/2018/09/a-compendium-of-conundrums.html

Logic puzzles

None of the puzzles below have trick answers - they can all be solved using logic and a bit of maths. Whenever a group of people need to achieve a task, assume they're allowed to confer and come up with a strategy beforehand. They're listed roughly in order of difficulty. Let me know of any other good ones you find!

Two ropes

I have two ropes which each, if lighted at one end, takes 1 hour to burn all the way to the other end. However, they burn at variable rates (e.g. the first might take 55 minutes to burn 1/4 of the way, then 5 minutes to burn all the rest; the second might be the opposite). How do I use them to time 45 minutes?

25 horses

I have 25 horses, and am trying to find the 3 fastest. I have no timer, but can race 5 at a time against each other; I know that a faster horse will always beat a slower horse. How many races do I need to find the 3 fastest, in order?

Monty hall problem (explanation taken from here)

The set of Monty Hall's game show Let's Make a Deal has three closed doors. Behind one of these doors is a car; behind the other two are goats. The contestant does not know where the car is, but Monty Hall does. The contestant picks a door and Monty opens one of the remaining doors, one he knows doesn't hide the car. If the contestant has already chosen the correct door, Monty is equally likely to open either of the two remaining doors. After Monty has shown a goat behind the door that he opens, the contestant is always given the option to switch doors. Is it advantageous to do so, or disadvantageous, or does it make no difference?

Four-way duel

A, B, C and D are in a duel. In turn (starting with A) they each choose one person to shoot at, until all but one have been eliminated. They hit their chosen target 0%, 33%, 66% and 100% of the time, respectively. A goes first, and of course misses. It's now B's turn. Who should B aim at, to maximise their probability of winning?

Duck in pond

A duck is in a circular pond with a menacing cat outside. The cat runs four times as fast as the duck can swim, and always runs around the edge of the pond in whichever direction will bring it closest to the duck, but cannot enter the water. As soon as the duck reaches the shore it can fly away, unless the cat is already right there. Can the duck escape?

Non-transitive dice

Say that a die A beats another die B if, when both rolled, the number on A is greater than the number on B more than 50% of the time. Is it possible to design three dice A, B and C such that A beats B, B beats C and C beats A?

Wine tasting

A king has 100 bottles of wine, exactly one of which is poisoned. He decides to figure out which it is by feeding the wines to some of his servants, and seeing which ones drop dead. He wants to find out before the poisoner has a chance to get away, and so he doesn't have enough time to do this sequentially - instead he plans to give each servant some combination of the wines tonight, and see which are still alive tomorrow morning.

a) How many servants does he need?

b) Suppose he had 100 servants - then how many wines could he test?

Crawling on the planet's face

Two people are dropped at random places on a featureless spherical planet (by featureless I also mean that there are no privileged locations like poles). Assume that each person can leave messages which the other might stumble across if they come close enough (within a certain fixed distance).

a) How can they find each other for certain?

b) How can they find each other in an amount of time which scales linearly with the planet's radius?

Dropping coconuts

I have two identical coconuts, and am in a 100-floor building; I want to figure out the highest floor I can drop them from without them breaking. Assume that the coconuts aren't damaged at all by repeated drops from below that floor - but once one is broken, I can't use it again.

a) What's the smallest number of drops I need, in the worst case, to figure that out?

b) Can you figure out an equation for the general case, in terms of number of coconuts and number of floors?

Pirate treasure

There are 5 pirates dividing up 100 gold coins in the following manner. The most senior pirate proposes a division (e.g. "99 for me, 1 for the next pirate, none for the rest of you"). All pirates then vote on this division. If a majority vote no, then the most senior pirate is thrown overboard, and the next most senior pirate proposes a division. Otherwise (including in the case of ties) the coins are split up as proposed. Each pirate's priorities are firstly to stay alive, secondly to get as much gold as possible, and thirdly to throw as many other pirates overboard as possible (they only pay attention to later priorities when all earlier priorities are tied). They have common knowledge of each other's perfect rationality.

a) What will the most senior pirate propose?

b) What about if there are 205 pirates?

c) Can you figure out a solution for the general case, in terms of number of coins and number of pirates?

Self-sorting

There are n people, all wearing black or white hats. Each can see everyone else's hat colour, but not their own. They have to sort themselves into a line with all the white hats on one end and all the black hats on the other, but are not allowed to communicate about hat colours in any way. How can they do it?

Knights and knaves

You are travelling along a road and come to a fork, where a guardian stands in front of each path. A sign tells you that one guardian only speaks the truth, and one only speaks lies; also, one road goes to Heaven, and ones goes to Hell. You are able to ask yes/no questions (each directed to only one of the guardians) to figure out which is which.

a) Can you figure it out using two questions?

b) How about one?

What is the name of this god? (explanation taken from here)

Three gods A, B, and C are called, in no particular order, True, False, and Random. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely random matter. Your task is to determine the identities of A, B, and C by asking three yes/no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for yes and no are da and ja, in some order. You do not know which word means which.

A game of greed

You have a pile of n chips, and play the following two-player game. The first player takes some chips, but not all of them. After that players alternate taking chips; the only rule is that you cannot take more than the previous player did. The person who takes the last chip wins. Is it the first player or the second player who has a winning strategy, and what is it?

Heat-seeking missiles

Four heat-seeking missiles are placed at the corners of a square with side length 1. Each of them flies directly towards the missile on its left at a constant speed. How far does each travel before collision? (Assume they're ideal points which only "collide" when right on top of each other).

Blind maze

You're located within a finite square maze. You do not know how large it is, where you are, or where the walls or exit are. At each step you can move left, right, up or down; if there's a wall in the given direction, then you don't go anywhere (but you don't get any feedback telling you that you bumped into it). Is there a sequence of steps you can take to ensure that you will eventually find the exit?

Hats in lines

There are 100 prisoners in a line, facing forwards. Each is wearing a black or white hat, and can see the hat colour of everyone in front of them, but not their own or that of anyone behind them; also, they don't know the total number of hats of each colour. Starting from the back of the line, each person is allowed to say either "black" or "white", and is set free if they correctly say the colour of their hat, but shot otherwise. Everyone in the line can hear every answer, and whether or not they were shot afterwards.

a) How many people can be saved for certain, and using what strategy?

b) (Very difficult, requires the axiom of choice). Suppose that the number of prisoners is countably infinite (i.e. in correspondence with the natural numbers, with number 1 being at the back). How can they save all but one?

c) Suppose that the number of prisoners is countably infinite, and none of them can hear the answers of the other prisoners. How can they save all but finitely many?

Prisoners and hats

Seven prisoners are given the chance to be set free tomorrow. An executioner will put a hat on each prisoner's head. Each hat can be one of the seven colors of the rainbow and the hat colors are assigned completely at the executioner's discretion. Every prisoner can see the hat colors of the other six prisoners, but not his own. They cannot communicate with others in any form, or else they are immediately executed. Then each prisoner writes down his guess of his own hat color. If at least one prisoner correctly guesses the color of his hat, they all will be set free immediately; otherwise they will be executed. Is there a strategy that they can use which guarantees that they will be set free?

Prisoners and switch

There are 100 immortal prisoners in solitary confinement, whose warden decides to play a game with them. Each day, one will be chosen at random and taken into an empty room with a switch on the wall. The switch can be in the up position or the down position, but isn't connected to anything. The prisoner is allowed to change the switch position if they want, and is then taken back to their cell; the switch will then remained unchanged until the next prisoner comes in. The other prisoners don't know who is chosen each day, and cannot communicate in any other way.

At any point, any prisoner can declare to the warden "I know that every single prisoner has been in this room already". If they are correct, all the prisoners will be set free; if not, they will all be executed.

a) What's a strategy that's guaranteed to work?

b) Does it still work if the warden is allowed to take prisoners into the room as often as he wants, without the other prisoners knowing? If not, find one that does.

Prisoners and boxes

Another 100 prisoners are in another game. They are each given a piece of paper on which they can write whatever they like. The papers are then taken by the warden, shuffled, and placed into boxes labelled 1 to 100 (one per box). One by one, each prisoner will be taken into the room with the boxes, and must find their own piece of paper by opening at most 50 boxes. If they do so, they're set free. To make things easier for them, before anyone else goes inside, the warden allows one prisoner to look inside all the boxes and, if they choose, to swap the contents of any two boxes (the other prisoners aren't allowed to move anything). Find the strategy which saves the greatest number of prisoners for certain.

Blue eyes (explanation taken from here)

A group of people with assorted eye colors live on an island. They are all perfect logicians -- if a conclusion can be logically deduced, they will do it instantly. No one knows the color of their own eyes. Every night at midnight, a ferry stops at the island. Any islanders who have figured out the color of their own eyes then leave the island, and the rest stay. Everyone can see everyone else at all times and keeps a count of the number of people they see with each eye color (excluding themselves), but they cannot otherwise communicate. Everyone on the island knows all the rules in this paragraph.

On this island there are 100 blue-eyed people, 100 brown-eyed people, and the Guru (she happens to have green eyes). So any given blue-eyed person can see 100 people with brown eyes and 99 people with blue eyes (and one with green), but that does not tell him his own eye color; as far as he knows the totals could be 101 brown and 99 blue. Or 100 brown, 99 blue, and he could have red eyes.

The Guru is allowed to speak once (let's say at noon), on one day in all their endless years on the island. Standing before the islanders, she says the following:

"I can see someone who has blue eyes."

Who leaves the island, and on what night?

Quine

Can you write a quine: a program that, when executed, prints its own source code?

Cheating on a string theory exam (puzzle taken from here)

You have to take a 90-minute string theory exam consisting of 23 true-false questions, but unfortunately you know absolutely nothing about the subject. You have a friend who will be writing the exam at the same time as you, is able to answer all of the questions in a fraction of the allotted time, and is willing to help you cheat — but the proctors are alert and will throw you out if they suspect him of communicating any information to you. You and your friend have watches which are synchronized to the second, and the proctors are used to him often finishing exams quickly and won't be suspicious if he leaves early.

a) What is the largest value N such that you can guarantee that you answer at least N out of the 23 questions correctly?

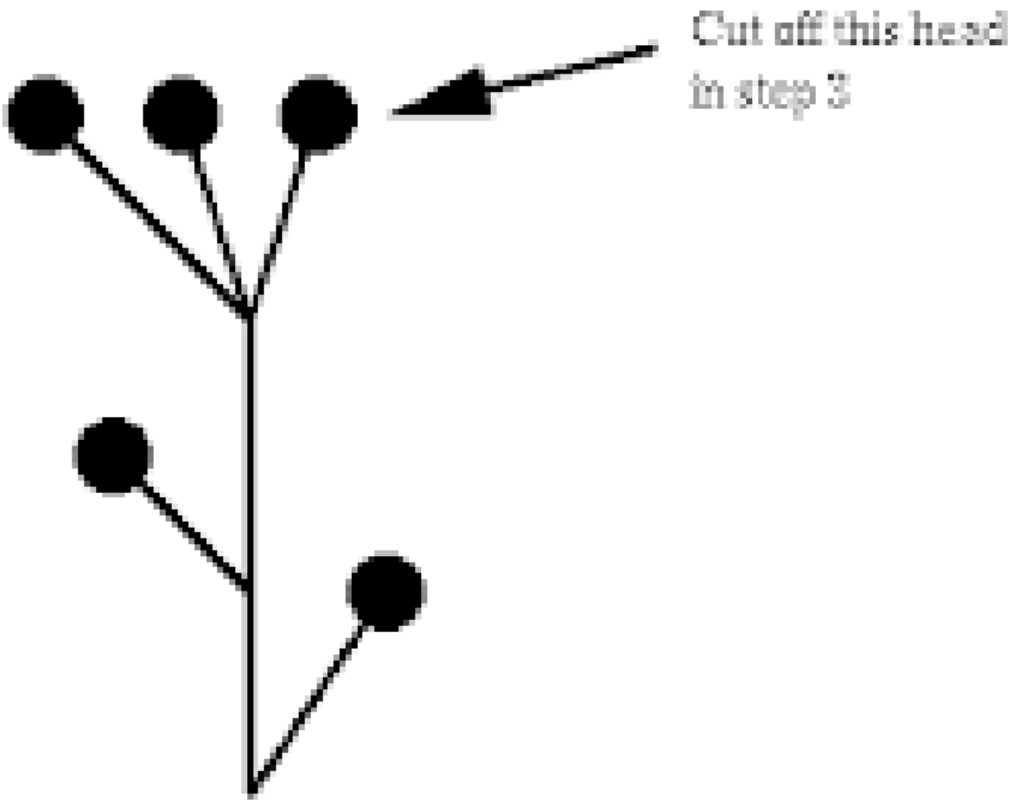
b) (Easier). The obvious answer is 12, but in fact you can do better than that, even though it seems like 12 is the information-theoretic limit. How come?

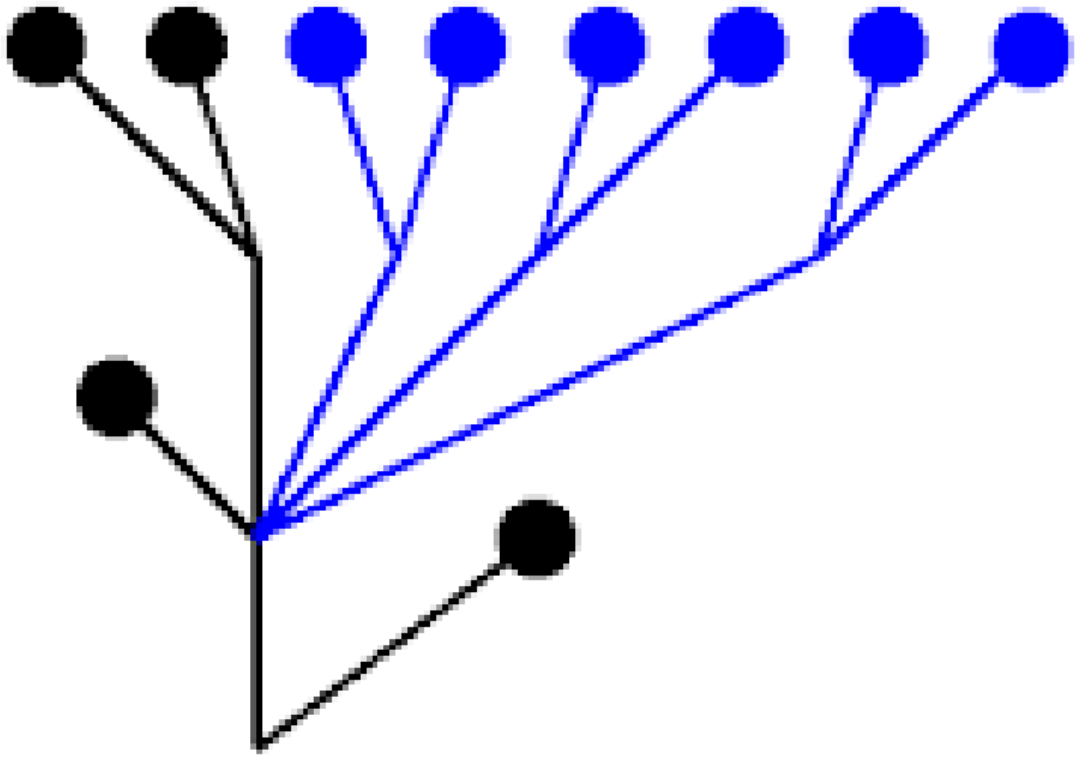
The hydra game (explanation taken from here)

A hydra is a finite tree, with a root at the bottom. The object of the game is to cut down the hydra to its root. At each step, you can cut off one of the heads, after which the hydra grows new heads according to the following rules:

- If you cut off a head growing out of the root, the hydra does not grow any new heads.
- Otherwise, remove that head and then make n copies of its grandfather subtree (as in the diagram below), where n is the number of the step you're on

What strategy can you use to eventually kill the hydra?


Cut off this head
in step 3



Physical puzzles

Balancing nails

Picture a nail hammered vertically into the floor (with most of it still sticking out). You're trying to balance as many other nails on it as you can, such that none of them touch the ground. How do you do so?

Hanging pictures

Consider a picture hanging by a string draped over some nails in the wall, in a way such that if any single nail is removed, the picture will fall to the ground.

a) Is it possible for 2 nails?

b) How about n nails?

Two-piece pyramid

Consider the two identical shapes shown below. Each has two planes of symmetry, and a square base. Is it possible to put them together to create a regular pyramid? (For a fun discussion of this problem in the context of machine learning, see a few minutes into this video).

Plane on a treadmill

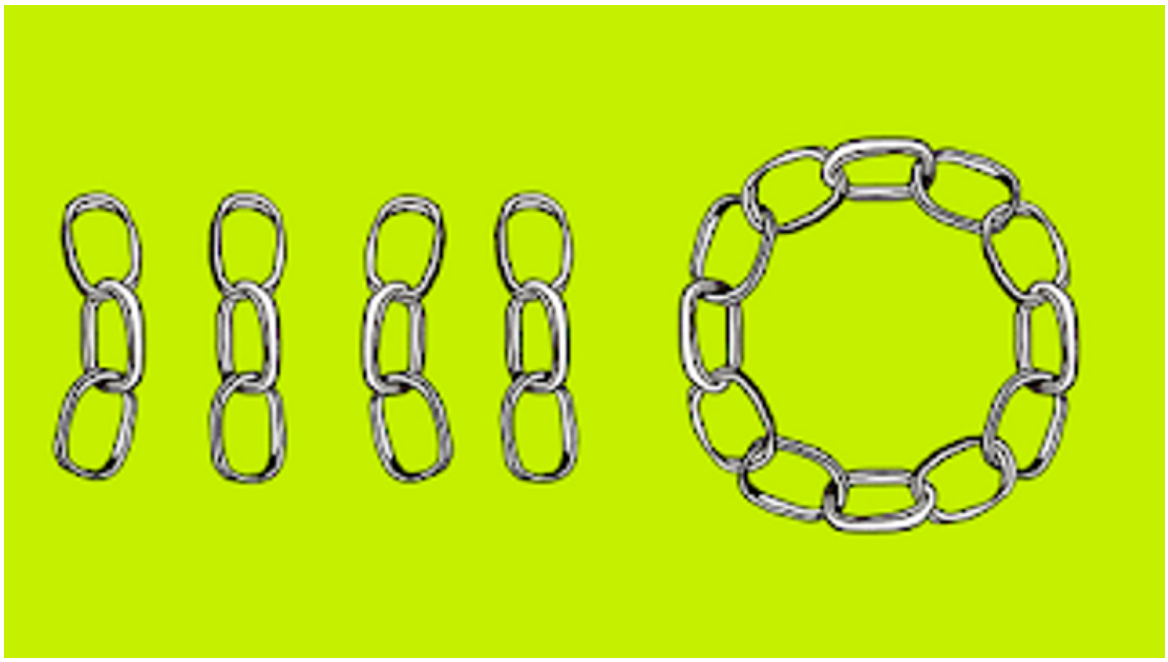
Suppose that a plane were on a gigantic treadmill, which was programmed to roll backwards just as fast as the plane was moving forwards. Could the plane ever take off?

Pennies game

Two players take turns to place pennies flat on a circular table. The first one who can't place a penny loses. Is it the first or the second player who has a winning strategy?

Joining chains

You have four chains, each consisting of three rings. You're able to cut individual rings open and later weld them closed again. How many cuts do you need to make to form one twelve-ring bracelet?

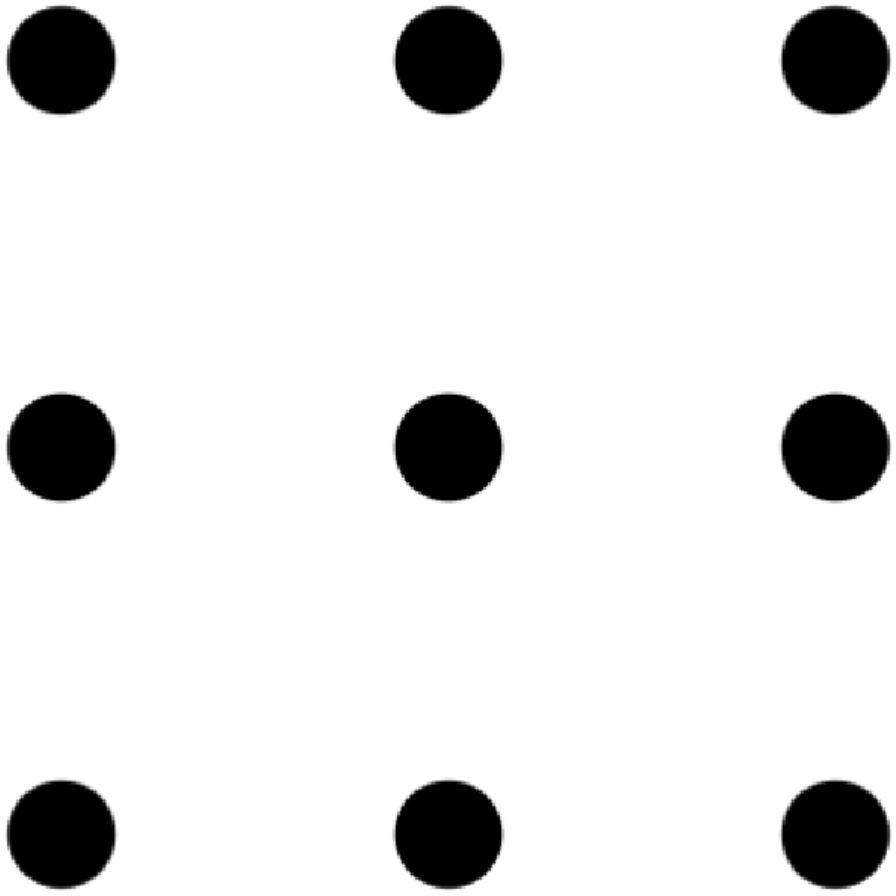


Going postal

Alice and Bob live far apart, but are getting married and want to send each other engagement rings. However, they live in Russia, where all valuable items sent by post are stolen unless they're in a locked box. They each have boxes and locks, but no key for the other person's lock. How do they get the rings to each other?

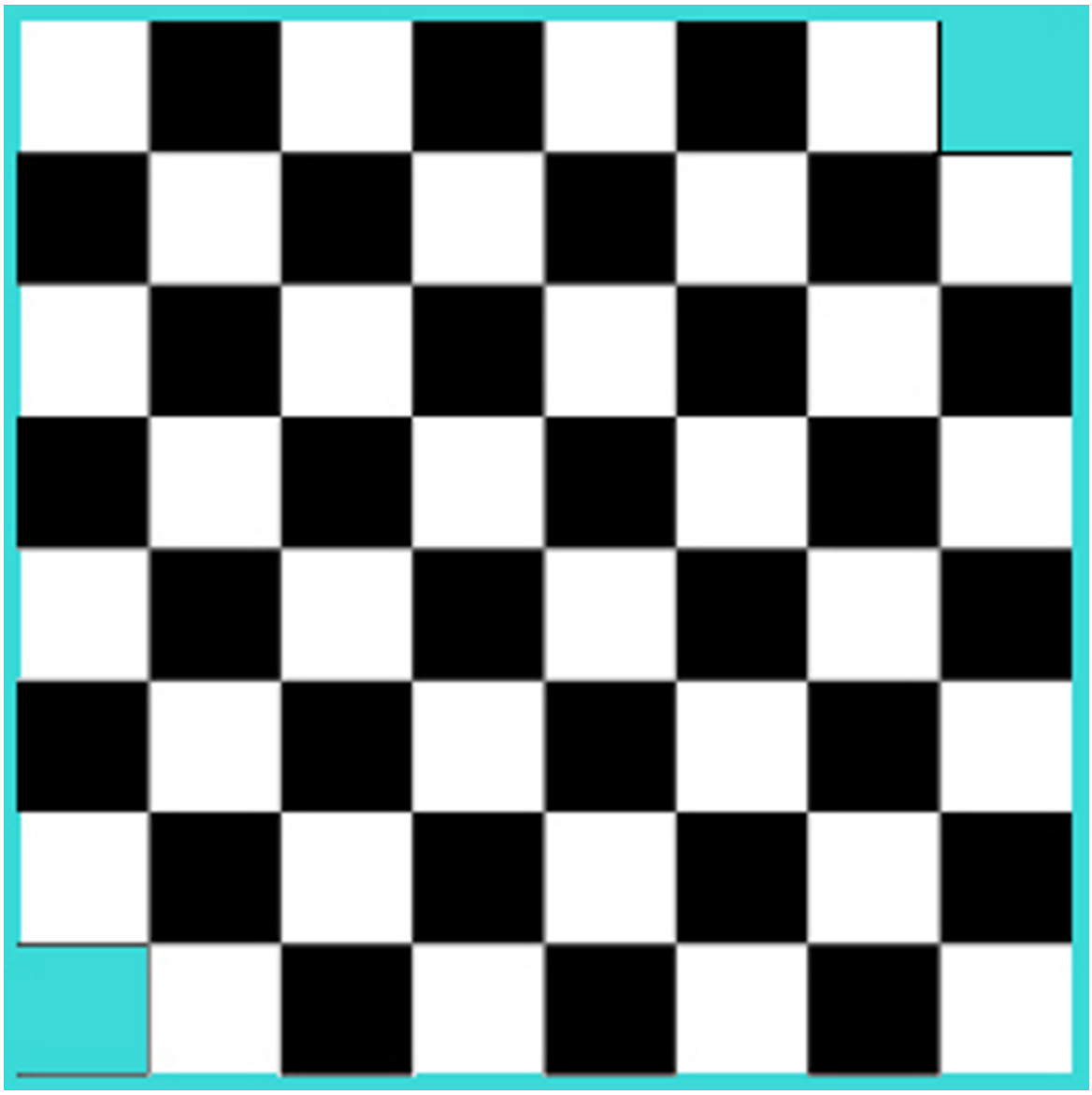
Nine dots puzzle

Without lifting your pen from the paper, draw four straight lines that go through the centres of all 9 dots.



Mutilated chessboard

Consider a chessboard missing two diagonally opposite corner squares. Is it possible to cover all the remaining squares with dominos (where each domino covers two adjacent squares)?

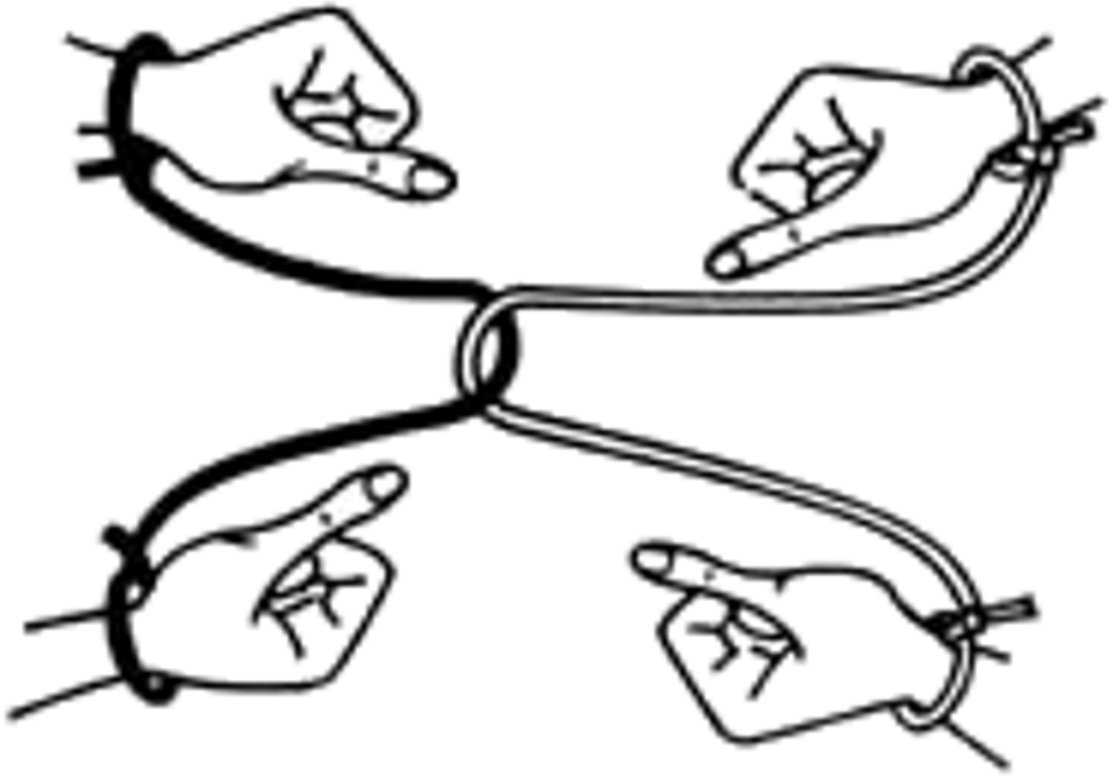


Safe sex

Suppose a man wants to have safe sex with three women, but only has two condoms. How can he do so, while ensuring that no STD is passed from anyone to anyone else?

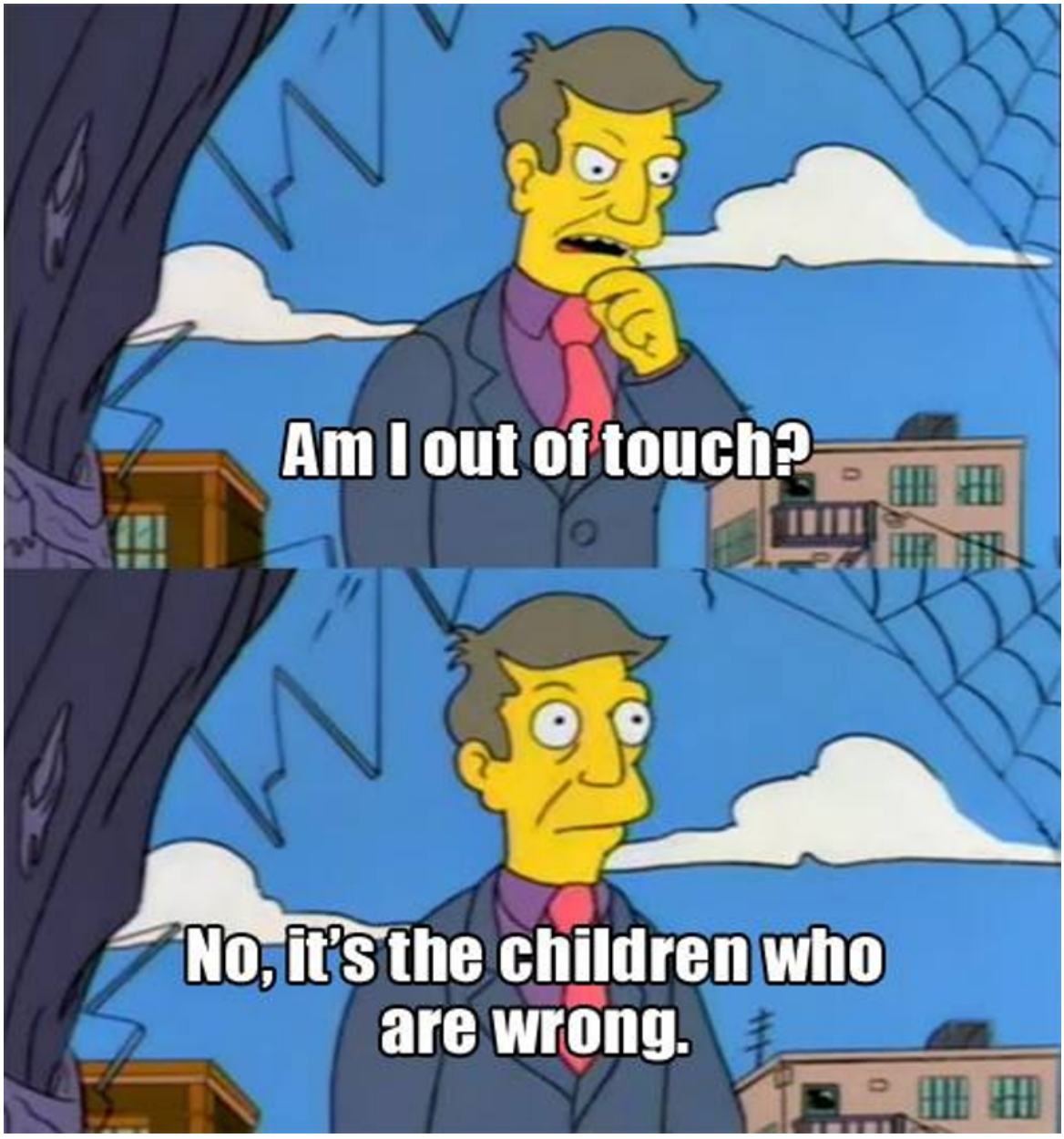
Wristcuffs

Two people are tied together as in the following diagram. Without being able to undo or cut the ropes, how can they get free?



Feedback from emotions

Bearlamp: Previous, First
Greaterwrong: Previous, First
Lesswrong: Previous, First



Mental health can often feel like the inability to get clarity around if it's, "just me" or if it's "the world" that's crazy. There's an open question in any interpersonal problem "is it me or is it them". Basic game theory might have you look at the general strategies and take a precommitment, like Tit for tat, with forgiveness. Something like, "It's always me" or "it's always them" - as the opinion that is formed in response to the stimuli being presented. These strategies tend to look like mental health problems when applied far too liberally. Some examples of these are in the List of Maladaptive Schemas.

If you play fixed mindset belief games, you will be bested by people who can see your fixed mindset and predict it. And beat it.

Unfortunately for *basic* game theory, *advanced* game theory comes along and sees all the other people playing with *Tit-for-tat, with forgiveness* strategies and generates a one-up strategy whereby advanced game theoretic players can beat basic game theoretic players, Just by playing one move ahead of the basic players.

(movie: The Princess Bride)

Unfortunately for advanced game theory, there exists expert game theory players who have seen that strategy and devised advanced strategies for solving the “how do I beat basic, and advanced game players”.

And unfortunately for expert game theory players there exists the halting problem. Where there will always be another level of play strategy. And there will always be another strategy taking into account all previous strategies. And this is an infinite loop.

how do I get feedback on an infinitely recursive system with the halting problem?

This question strikes at the core of the interface between self and the external world. We are each a chinese room brain. This is the problem of other minds. When we design an experimental apparatus and attempt to glean feedback information from reality as if we are not in it, we don’t really answer the question here.

I only have one answer. And it’s an unfortunately frustrating one. I hint at the answer in the emotional training model but that’s not ultimately obvious enough.

Feedback has to come from within.

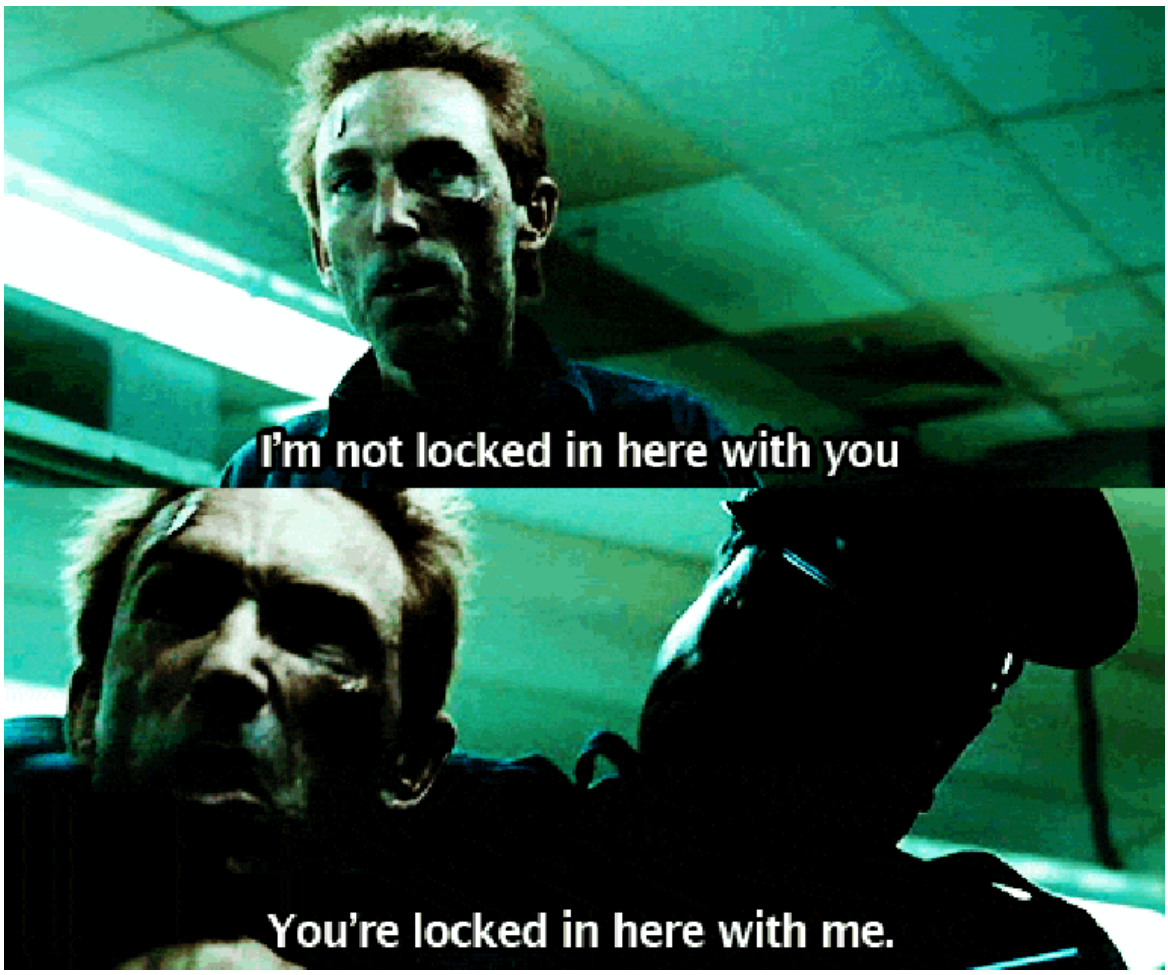
How do I know what to do? How do I gauge what is right and wrong where all I have to go on is the intention to gage right and wrong, and a collection of informational experiences that form my sensate reality including knowledge I have gathered by reading books, talking to people and experiencing life myself?

There is no “truth grain” external to the self; where, having found the truth grain, there is no need to be wrong ever again. There is no fundamental reason why we can believe and trust external information more than internal information. (external information is only internally represented after all - with an assumption that we can comparably across brains; form equivalent internal representations of external information.)

I am ae enclosed brain. Feedback has to come from within the system. When I look in a mirror, I see a reflection of myself, but the reflection registers in the system. The results of the reflection “wow I like the way I look” is a judgement call that happens from within the system. When I ask my friend how I look and I receive the information that “I look as ugly as a bat out of hell”, that information registers inside the system. Inside the brain. External validation is an illusion.

In that sense, if I didn’t already, now would be a good time to start liking myself.

Because…



Next: The third system

Effective Altruism Book Review: Radical Abundance (Nanotechnology)

Book Review: Radical Abundance

I. Introduction

As a materials engineering major with, roughly speaking, a year of full-time experience with molecular dynamics simulations, I have a special place in my heart for high impact materials both literal and figurative. As a lurker in effective altruism, I was delighted to see something potentially relevant to my experience show up. I made a post in the "Nanotechnolgy in EA" Facebook group expressing/re-iterating the following thoughts:

1. Nanotechnology will be chaotic and hard-to-predict, requiring massive advancements in computation before effective machines can be designed.
2. Our technology in terms of actually producing APM seems plausible but very uncertain at this point

From my post, I got some interesting comments suggesting an alternative to 1. along with a recommendation to read K. Eric Drexler's new book on atomically precise manufacturing (APM), *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. This, in part with some googling, gave me the impression that I was also wrong about 2.

The goals of this book report from least to most significance are: providing a decent summary of *Radical Abundance,* expressing my own thoughts on what I find to be the most salient aspects of nanotechnology, and clarifying possible common misconceptions that contribute to the frequent confusion about nanotechnology's place in effective altruism.

Anyway, Radical Abundance lightly discusses three inter-woven concepts: how APM would work, how APM could impact the world, and how APM can be achieved. Let's do this.

II. How APM Works

A. Advanced APM in a Nutshell

The description and understanding of APM can be best characterized by Drexler's own words:

APM-based materials processing technology will employ nanoscale mechanical devices that operate at high frequencies and produce patterns of atoms...think of an APM system as a kind of printer that builds objects out of atoms just as a printer builds images out of patterns of ink, constrained by a limited gamut, not of colors, but of output materials.

This clarifies two distinct points that I, as someone working in molecular dynamics and , would miss. First, despite my priors from experience with soft matter molecular dynamics, APM systems will be based on mechanically-inspired rigid mechanisms rather than biologically-inspired soft matter machines. Second, at its heart, most APM

will focus on the fabrication of macro-scale materials with nano-scale optimization (i.e. really good materials) rather than nanomachines (i.e. tiny robots), though these will be necessary to some extent in creating APM in the first place and have massive peripheral uses and implications.

B. Mechanical Devices in Nanotechnology

First, while existing examples of APM (i.e. ribosomes) are biological in nature and demonstrate the potential for leveraging nano-scale devices in complex ways, Drexler's APM uses down-scaled mechanical devices like those in modern factories to achieve its ends instead of the chaotic complexity of biology. In order to successfully imitate the behavior of mechanical machines, nanomechanical machines will typically be made of "stable covalent structures that consist of fused rings; among hydrocarbons, small-scale examples include the adamantanes and the somewhat more flexible aromatic molecules." Thinking along these lines simultaneously opens up new possibilities and grounds others. When I think about biological APM, I think of spectacularly complex inter-plays of atoms and molecules superbly optimized in parallel with their environments over millions of years to provide the sufficient but incredibly improbable conditions for life. When I think about mechanical APM, I think about the Saturn V Rocket.

To illustrate the difference, consider the following situation where someone asks me about the future of leveraging nanotechnology to achieve goal a goal in a given environment:

Me thinking about biological nanotechnology: "Of course! We'll use incredibly complex computational tools coupled with automated laboratories to build a black-box machine that does the job you need… At least, we will if we ever figure out how to make computers and laboratories that effective. Once we do though, the sky's the limit!"

Me thinking about mechanical nanotechnology: "Hmm... It looks like achieving your goal also requires some flexible components or at least a fair number of automated computerized systems so I'll need to see if we can get materials to work well with that at the nano-scale… And your environment is a bit different so I might need to select different molecules to make sure they're stable and rigid unless you want to incorporate homeostasis which would add another order of complexity... Give me five years, a few millions dollars and the atomically precise tech I need to build something and I'll let you know."

The difference between these biological and mechanical characterizations of nanotechnology is simultaneously complicated and subtle. Both seem a long ways off in terms of technology but the former demands a clear conceptual leap (much better computers) while the latter requires an extensive but plausible level of technological progress (Improved knowledge of chemistry and existing chemistry techniques). This means that, while the former might never be achieved, the latter will probably be achieved eventually.

A similar point can also be made about potential impact. Biological interpretations of nanotechnology very quickly start hinting at massive revolutions in nano-machines that constantly respond and improve all aspects of human life and engineering, aggregating together to form macro-level machines of incredible complexity and computational capacity. In contrast, mechanical interpretations of nanotechnology are grounded in limited types of materials in limited environments with limited capabilities.

For the most part, these limitations are the same for both macro-scale mechanical devices and nano-machines since both can be described in "familiar, mechanical terms." However important differences do exist. Drexler notes that "a machine can't work well if its parts can't move smoothly and interactions between atomic-scale bumps on surfaces might seem to make smooth movement impossible." This can be resolved through designing bump patterns to produce superlubricity but comes with its own constraints. Drexler also notes the presence of drag and thermal motion but these forces are not very significant in the context of rigid mechanical structures. I would also like to add an additional constraint: Because "stretching space and time in equal proportion scales properties like mass, force, and velocity in exactly the right way to make mechanical motion the same," one quickly realizes that gravity, producing an acceleration of 9.8 meters/second down regardless of size, vanishes in importance. In this sense, nano-scale factories will not be like the factories seen on Earth but rather more like the ones designed in space. Additionally, they will be driven purely by motors and mechanical devices with no incorporation of electrical wiring (classical electrical engineering breaks down at the nano-scale).

Initially, these limitations made me skeptical about whether APM could feasibly create complex designer materials or extend to advanced nano-machines outside of a manufacturing context. However, this video led me to realize that the sort of target materials to be constructed could rely on basic patterns of picking up and putting down different atoms and molecules. Additionally, because the nano-scale is so small, relatively large nano-machines can still be build at very low size-scales (though this nano-machines are still what I am most skeptical about).

For the most part though, the outputs of my biologically-inspired expectations of nanotechnology and Drexler's mechanical vision seem similar. As Drexler says:

A SCIENTIST WROTE an article about the nanomachines of the general sot I've described, but he suggested that they couldn't be used in a biological environment because biomolecules would gum up gears and other moving parts. The answer, of course, is to keep gears in a gearbox, and to place all critical moving parts inside a sealed shell.

The moral here is that most mechanical nano-machines can be designed to avoid many of the problems that they might face relative to their magical biologically-inspired cousins. At the same time, this indicates a more general problem: to build a functional nano-machine, there must be some input and some output which facilitates action on the environment. This means that some of the mechanisms in a given nano-machine must be designed for the environment. However, the severity of this limitation also depends on the nano-machine's functionality: For power, mechanical mechanisms can be designed to minimize complex parts exterior to a core gearbox. For expelling objects, a nano-machine might incorporate a simple airlock-like design. However, for taking in and processing new molecules in an uncontrolled environment, nano-machines would require both filters and safe-guards to prevent clogging.

C. APM as Macro-Scale Manufacturing

While these limitations do not preclude the existence of nano-machines (and especially those that manufacture on the fly) in uncontrolled environments, they do raise questions. However, this specific technology is not the end goal that Drexler focuses on in *Radical Abundance*. Though Drexler does allude to "fast, thorough data collection and the means for rapid deployment of nano-scale devices" in the context

of biological interventions, the lion-share of speculation about the future is focused on machines like this one:

Picture yourself standing outside the final assembly chamber of a large-product APM system and looking in through a window to view the machines at work in a space the size of a one-car garage....

To the right, you see an exit door for products ready for delivery. To the left, you see what look like wall-to-wall, floor-to-ceiling shelves, with each shelf partitioned to make a row of box-shaped chambers. In the middle of the garage-sized chamber in front of you is a movable lift surrounded by a set of machines.

The machines look uncommonly sleek, yet very familiar. They resemble machines in an automated factory, with robotic arms programmed to swing around, pick up components, and swing back to snap the components together. The machines look like this because they are, in fact, machines in an automated factory and because machines that perform similar motions often have similar shapes and similar moving parts. Because they are made of materials better than steel, however, they can be faster, lighter, and more efficient.

Looking back at the wall on the left, you can get a clear view into several chambers that happen to be at eye level and near the window. Each smaller chamber contains machines with swinging arms, and the overall setup inside looks like a scale model of the larger chamber, complete with a rear wall with wall-to-wall, top-to-bottom rows of yet smaller chambers. It’s hard to see in detail what these small chambers-within-chambers contain, but they seem to hold a tiny yet familiar set of machines mounted in front of a rear wall with rows of yet smaller chambers.

With the press of a button, the machinery kicks into gear. At first nothing seems to happen, but in less than a minute the large machines in front of you start to pick up parts as they pop out of the chambers in the wall at the left, moving these parts to the platform in the center where the first parts are clamped, and the rest snap together. As the machines put the parts together, a familiar product takes shape, an automobile, different in almost every detail from those built today, yet having a form that reveals the same function. chambers.

With the press of a button, the machinery kicks into gear.

At first nothing seems to happen, but in less than a minute the large machines in front of you start to pick up parts as they pop out of the chambers in the wall at the left, moving these parts to the platform in the center where the first parts are clamped, and the rest snap together. As the machines put the parts together, a familiar product takes shape, an automobile, different in almost every detail from those built today, yet having a form that reveals the same function.

Each part takes several seconds to put into place and new parts slide out of the chambers at a corresponding rate, each chamber delivering a component every few seconds. To the left, inside the closest chamber, you can see the machines working inside. These miniature machines seem to be performing similar tasks, but at a rate of several cycles per second, their motions are almost too quick to follow. It’s easy to guess what’s happening in the yet-smaller chambers farther back, yet the motions there are no more than a blur.

In the main chamber the work is complete in less than a minute. The door to the right then unseals and opens, and a car moves out into a receiving area, sealed in what

looks like a plastic sleeve. A moment after the door reseals, the sleeve is pulled back for recycling and the process is done. (This exit maneuver is part of a cycle that prevents contaminants from entering when the product exits.

This description is in line with the video linked earlier and leads me to infer that, though many other uses also exist for nanotechnology, the main use-case emphasized in *Radical Abundance* is extremely efficient factories that build extremely high quality materials by exploiting "lightweight, carbon-based materials" and interesting already-discovered patterns that may produce exotic electronic properties on the macro-scale. I mention this because APM as advanced manufacturing and APM nano-machine technology itself reflect different risk profiles that ought to be discussed in the impact section. As Drexler says, "where the physical nature of APM technologies is concerned, the relevant questions pertain to the physics and engineering of compact, highly capable factories-- not vague dreams, exotic products or nanobugs."

III. APM Will Change Everything

A. Drexler's Partial Notes on APM Benefits

Because APM will use cheap, readily available materials like carbon to produce high quality materials in minutes from a compact machine, it has the ingredients to revolutionize the existing means of production:

"Nanoscale size enables extreme productivity as a consequence of mechanical scaling laws. In addition, small-scale, versatile, highly productive machinery can collapse globe-spanning industrial supply chains to just a few links."

This massive simplification of supply chains coupled with APM's use of abundant molecules implies a massive economic shift. Locations impoverished by lack of capital, business connections and resources could, in principle, use nanotechnology to sustainably and locally produce high quality infrastructure and reliable food sources.

Other benefits of APM include transforming information technologies through more advanced materials, improving infrastructure through better construction and transportation technology, improving agriculture through efficient recycling and water-processing/desalination, and resolving global warming carbon dioxide capture. The running theme themes in each of these solutions are APM’s blend of efficiency and ability to cheaply construct many of the advanced nanomaterial solutions which currently exist but are too costly to implement industrially.

The benefits of APM in improving agriculture also struck me as particularly significant. In particular, Drexler notes that enclosed agriculture (i.e. large-scale greenhouses) offers "higher yield per hectare, better food quality, freedom from pesticides, extended growing seasons, freedom from constraints of soil quality and available water, and protection from drought" while also reducing "water demand and contamination." These latter benefits—freedom from soil, water and weather constraints—are already achieved achieved in modern greehouses and act as the corner stone for alleviating starvation in impoverished regions. Nevertheless, the modern world, let alone impoverished regions, cannot meet the efficiency and infrastructural requirements that would make implementing greenhouses economical. APM’s capacity for rapid and cheap infrastructure along with highly efficient energy systems (i.e. solar power) will change that.

Overall, I see only two reasons to doubt these benefits. The first reason is economic. Historically, international trade has been a necessity for technologically advanced

countries which has allowed the flow of money to less developed countries. However, APM will finally allow technologically advanced countries to become self-sufficient to the extent that they do not need to import goods from less advanced countries. If these countries cannot amass the wealth to gain nanotechnology, the absence of economic ties with technologically advanced autarkies may lock them into poverty. However, the presence of these self-sufficient countries also makes this situation unlikely. Because APM-driven countries will be so wealthy and likely retain some altruistic leaning, the only reason for them not to provide APM to less advanced countries would be some sort of non-economic cost or risk—that is, a military risk. Fortunately, because APM factories will likely be both technologically complex and optimized for use in fabricating specific materials rather than every material, the military risk of providing factories designed with food and infrastructure in mind is relatively low. This is especially true when noting how military APM will empower the most advanced countries relative to others. To be fair though—in the era of semi-costly/only partially effective nanotechnology—economic concerns could be very serious even if they are only short-term considerations.

The second reason for doubt is more significant and relates to an important criticism of *Radical Abundance*: Drexler effectively discusses the capabilities of nanotechnology but does not seriously discuss the details of the problems he claims that it will solve. In practical terms, the most important altruistic contribution of nanotechnology would be poverty reduction. Explicitly, nanotechnology should provide sustainable food sources in places where food cannot currently grow, water in places where water is currently inaccessible, infrastructure in places where infrastructure cannot currently be established, and energy in places with limited energy resources. These problems are almost all geographical constraints and, while Drexler claims that "the most useful elements--including carbon nitrogen, oxygen, and silicon--are not all that scarce," nitrogen deficiency is one of the reasons for agricultural difficulties in Africa. Moreover, many impoverished countries suffer from lack of water/reliance on contaminated ground water and, while nanotechnology may provide cheap purification methods, I am not convinced that it will either offer purification methods or water transportation methods better than the ongoing science in those fields. Finally, many impoverished countries suffer from rocky, mountainous and swampy terrain which inhibits movement and the feasibility of taking advantage of large-scale infrastructure. Enclosed agriculture on forty-five degree rocky inclines does not seem very promising, especially when compared to just using longer supply chains which APM would hopefully avoid. In short, while nanotechnology might allow the collapsing of globe-spanning supply chains to a few efficient links, Drexler does not discuss how the locations that are most harmed by the need for these supply structures would surmount their existing geographical constraints using nanotechnology. If this is not addressed, it creates a massive problem for APM. After all, if geographical issues limit the efficacy of altruism through nanotechnology, then—for impoverished countries—access to nanotechnology may not outweigh the economic loss of being unable to trade with wealthy countries.

Beyond the potential for providing the basic staples of human life, an adjacent significant contribution that highly advanced nanotechnology brings is control of the ecosystems that make these staples. As someone concerned about wild animal suffering, this development could massively reduce human harm to insects and potentially offer strategies for systemic wildlife interventions. While restraint and hesitation should generally be applied to actions like this as they currently have dramatic and unpredictable impacts on wild ecosystems, the extensive surveillance technologies provided by APM (discussed later as a drawback) along with advancing

science may provide the exact kind of ecological knowledge needed to make these adjustments wisely.

While I have previously suspected that limitations on chemicals and manufacturing methods may render APM less useful than it seems, I think that these views fail to appreciate simply how much can be accomplished through the cheap production of high quality infrastructure and electronics alone. With this in mind, I think that the only scenarios where APM is less-than-revolutionary are scenarios where some other easier-to-develop technology focused on particularly important use-cases succeeds faster by virtue of a more direct path to implementation. Off the top of my head, these might look like some of the following:

- Plant growth in barren environments through genetic engineering couples with mass automation of farming to produce cheap, abundant food that is not limited by distribution costs.
- Some large-scale technique for processing abundant elements efficiently into useful materials is established (though I am not aware of methods for this outside of nanotechnology)
- Mass trial and error facilitated by microfluidics enables sufficiently fast and personalized medical treatment that APM may only be able to enhance rather than revolutionize
- The gradual arc of science improving efficiency across the board in transport logistics coupled with existing information technology drive a shift to a digital, de-localized economy before APM makes universal wealth feasible.

This is my greatest and most explicit uncertainty with respect to APM's benefits. I agree that APM would provide these benefits but I am not sure whether APM will be the first to provide them. Heuristically, I usually expect that technology specifically designed to address a specific issue will advance much faster than a more generic abstract approach. This leads me to doubt the comparative benefits of pursuing APM as opposed to other more explicitly oriented research endeavors. On the other hand, APM is a relatively concrete idea and, even if it is likely more difficult to pursue than a more focused technology, may be worth it due to its large potential range of impact.

B. Drexler's Notes On How APM May Go Wrong

One of the dangerous aspects of APM that Drexler notes is surveillance.

Small-scale components and systems open new possibilities. For perspective, consider that a one-gram platform built with advanced technologies could produce teraflops of computational power (and much more, in bursts), together with a million-terabyte data storage capacity and better-than-human sensors, all with a power demand comparable to that of a cell phone on standby. Now consider what could be done if devices of this class cost about $1.00 per kilogram and could be delivered by small drone aircraft. $100 billion would buy roughly one device of this sort per square meter of land area, worldwide.

Because most of this technology will not need teraFLOPs of computing power, this raises the possibility of globally abundant third-party surveillance as easy to bring to work as it is to track dirt into a house. Coupled with the ability to relay information, these sensing devices would be able to trivially transport massive amounts of data to a centralized system for fast (>petaFLOP) processing. This could quickly lead to a race-to-the-bottom favoring those who are willing to care less. Combining these same devices with chemical payloads may also enable groups to engage in the soft

psychological manipulation or undetectable assassination of their enemies. Combining these "secretive surveillance regimes" with a world that replaces wealth with individual virtue as a status symbol in absence of poverty may also lead to the micro-regulation of human behavior. Additionally, because surveillance does confer advantages onto the groups that use it and many countries vary significantly in their attitudes toward individual privacy, establishing a common pro-privacy agreement to avoid a race-to-the-bottom will be unlikely.

Overall, I am not concerned about this risk mainly because this surveillance technology is oriented around active nano-machines in complex and unpredictable environmental situations. Unlike the benefits of nanotechnology which are produced in controlled manufacturing contexts, this technology must function in a less manageable space and I expect physical and design limitations to seriously come into play here. Undoubtedly, nanotechnology will cheapen surveillance but I doubt that it will be sufficiently cheap to subvert personal privacy more than it would be anyway. Furthermore, while a race-to-the-bottom may happen, it may be partially mitigated by the radically different economic circumstances which more simple nanotechnology will have already provided the world, making it "hard to find an external resource that would continue to be a vital national interest."

These are also the same reasons that I am not seriously concerned with nanotechnology causing massive environmental damage through self-replicating robots: the problem is uniquely challenging and strong incentives for doing so appear to be absent. In fact, Drexler notes that nations have "strong, shared interests in constraining non-state actors" from the use of unrestrained APM. A weaker variant of this issue is the potential for nanotechnology to create nanoparticle pollutants on a massive scale, leading to massive ecosystemic damage. However, countries capable of producing nanotechnology at this level will also likely have technology capable of rapidly identifying (via surveillance) and addressing pollutants anywhere on the globe. This means that mass pollution is only a risk if done by a country that would not fear retaliation. Because this type of country can already threaten the world in many more ways than just pollution, I consider this to be a minor concern.

The second main risk that Drexler discusses is a two-pronged change in military strategy. First, because APM will be used to design sequential generations of APM, "it is therefore easy to envision scenarios in which a modest degree of asynchrony, measured in months or less, could swiftly lead to radical military asymmetries even with relatively shallow exploitation of the overall potential of APM-level technologies." Second, APM would provide "abundant, affordable, non-lethal, remotely operated weapons" that reduce risk of harm and encourage more war-like policies. While these shifts may encourage technologically advanced countries to pursue aggressive expansionist policies, expansionist actions will offer little in terms of resource gain and cost a level of societal stability due to cultural conflicts. In fact, because of the absence of resource gain, expansionism will primarily be driven by a desire to establish different societal values--a goal that governments will more likely pursue through subtler methods than outright conflict.

Lastly, Drexler also notes that one of the commonly discussed risks--the idea that self-reproducing nano-machines will consume vast swathes of the world--is unlikely given a manufacturing model of nanotechnology instead of a biological model. I agree that this is unlikely for the same reason that I saw surveillance as unlikely but multiplied several-fold. Building environmentally robust reproducing nano-machines using abundant natural supplies and establishing enough stable complexity for those nano-

machines to reconstruct themselves both require an enormous amount of effort far beyond the scope of the primary benefits of nanotechnology which Drexler discusses.

One danger that I believe merits concern but goes unmentioned by Drexler pertains to advanced general intelligence. If Drexler is correct in his claim that nanotechnology will offer teraFLOP/gram computers, then an average laptop computer weighing five pounds would have about a petaFLOP of computational power and a thousand of such computers would start pushing on the capabilities of a human brain (according to Kurzweil) without any additional deliberate optimization beyond emulating nature. This is more than sufficient to produce intelligence far more competent than any human and intertwines nanotechnology deeply with its relationship to artificial intelligence. If humanity is not confident in its ability to align the goals of advanced general intelligence (which it is not), then nanotechnology in this regard ought only be pursued if does not increase the risk of misaligned advanced general intelligence. This will be the case if either advanced general intelligence is expected to emerge prior to advanced nanotechnology (in which case AI Safety research is a much more pressing concern than nanotechnology) or Drexler is incorrect about teraFLOP/gram computers.

Overall, my opinion is that Drexler's benefits of nanotechnology outweigh Drexler's risks but, when I consider nanotechnology's relationship with artificial intelligence, I begin to suspect that it is far more dangerous than it initially appears. That being said, this only applies directly to the development of nanotechnology itself. Development of nanotechnology *policy* on the other hand may have significant value in the case where humanity weathers advanced general intelligence. However, the benefits of this work may be limited by the likelihood that advanced general intelligence will transform the political landscape to such an extent that any current policy design would be inapplicable.

IV. Pathways to Nanotechnology

When wondering about the likelihood of achieving something like APM, a useful question is "Why has this not been achieved already?" This is an especially salient question for APM both because of the extensive funding providing in the field of nanotechnology and the relative age of the idea itself. Drexler answers this question with two main explanations. The first is that extensive funding may have been focused on "nanotechnology" but minimal funding has been focused on APM.

"The great promise of nanotechnology is atomically precise manufacturing, and the US Congress established a program directed toward this objective, but the program instead did something entirely different. The program's leaders redefined "nanotechnology and supported only nanoscale materials and devices, technologies as different from APM as cloth, cement, and wires are from a programmeable digital computer. Most research advertised as "nanotechnology" has therefore been irrelevant to what had been widely expected, and while atomically precise fabrication has flourished in the molecular sciences, people looking for progress toward APM-level technologies have been led to look in the wrong direction."

The second is a rather philosophical discussion about how the basic science that mostly comprises nanotechnology research is ineffective at producing the system-engineering coordination required to make an actual technology. As Drexler says, "no matter how research-intensive a project may be, work coordinated around concrete engineering objectives will eventually be required to produce concrete engineering results." So far, work at this level of coordination has not been pursued.

Both of these explanations demonstrate that the existing limited state of APM is due to political and organizational constraints that do not reflect negatively on the feasibility of APM itself. As for that feasibility, Drexler notes that *designing* the mechanisms of APM devices is already relatively feasible: "Machine components based on rigid covalent structures cannot yet be implemented, yet are already easy to design and model using standard computational chemistry software." Indeed, I suspect that one way to motivate APM-directed research would be to use the tools we have to preemptively design an advanced APM machine and then to use that device as a motivator for nations to funnel research into its ultimate construction.

The main current method for developing APM will involve advancing stereotactic chemical reactions. While "conventional, solution-phase chemical reactions are enabled by local structural features of molecules," more complex synthetic targets with many more structural features would be "difficult or impossible to direct reactions with sufficient specificity." Stereotactic chemical reactions address this issue by having "linking structures direct reactions by constraining encounters among potentially reactive groups, separating some pairs while increasing encounter rates between others." By definition, stereotactic chemical reactions reflect the atomically precise manufacturing of complex structures through sequentially bonding molecules together. This description applies to biological proteins as well but unlike proteins, APM will develop stereotactic reactions through mechanical designs rather than thermally driven ones.

Drexler notes that stereotactic reactions in advanced APM must satisfy a number of constraints: structures must be stable and rigid; structure motion must be well controlled; reactions must happen reliably; reactions must be irreversible; and reactions must yield only a single product. Many of these requirements (motion control, reliable processes, irreversibility, single product) are consistent with most fabrication procedures. However, rigidity holds special importance for APM because it allows fabricated materials to occupy meta-stable states. This means that, while flexible biological materials can quickly restructure themselves into lower energy/higher entropy states by just moving, APM materials move in a slow manner due to rigidity that prevents them from finding those states as quickly. This ensures that while APM materials, like diamond, maintain their properties for a long time even though they might not be naturally stable.

Current technology cannot meet all these requirements however the assembly of "polymeric building blocks, cross-linked via conventional reactions, with all stereotactic operations performed in aqueous environments" may readily be achieved through existing methods like DNA origami. This offers a bottom-up APM starting point wherein "large self-aligning building blocks, loose positional tolerance margins, low stiffness materials, simple machines, and simple motion constraints" are iterated upon to produce the more complex, more precise, and higher quality nano-devices needed to reach advanced APM. In other words, simply iterating on existing nanotechnology related to stereotactic control will facilitate the improvement of stereotactic control.

Drexler also addresses another aspect of nanotechnology that had previously led me to be more suspicious about its success: On the macro-scale level, a factory may rely on machines produced by other factories but those factories themselves are assembled by humans (though often humans piloting machines themselves). On the nano-scale though, human assembly is impossible so factories would need to produce factories--which seems hard. In actuality, the solution is simple and practical: Stereotactic control will lead to "improvements that facilitate the design and fabrication of complementary surfaces... Stereotactic synthesis can enable advances

in component level self-assembly." Once humanity can produce nano-machines, humanity will be able to produce nano-machines attached to surfaces with binding affinities for complementary surfaces. At this point, near current-day technology would be able to produce large-scale complementary surface patterned sheets that bind only the right nano-machines in only the right places. In other words, success will not initially be achieved through a closed system of nano-factories making nano-factories but rather through a blend of manufacturing nano-machines to more easily interact with higher level self-assembly methods.

V. Conclusion

Throughout *Radical Abundance*, Drexler emphasizes two common misrepresentations of nanotechnology: the popular representation of nanotechnology as almost magical nanobot swarms and the academic representation of nanotechnology as pertaining to science at a given size-scale rather than technology at a given size-scale. As someone both aware of popular culture and closely involved with conventional academic research in nanotechnology (i.e. self-assembling nano-particles), I blended these ideas together into an improbably influential technology mediated by an improbably hard-to-control science. In reality, APM is a comprehensibly impactful technology mediated by under-researched but high-potential iterative design. Finishing the book, I feel reasonably convinced that APM is achievable based on existing technological progress, partially convinced that APM will have the benefits that Drexler claims it has, and unconvinced that the benefits outweigh the risks if we limit them to those that Drexler has described.

Nevertheless, while understandable in the context of nanotechnology's previous over-hyped history in pop-culture, I think *Radical Abundance*'s discussion of risks is too conservative in estimating their magnitude. This is because, in a significant number of circumstances where human-developed nanotechnology becomes important, I envision scenarios that accelerate the emergence of a misaligned general artificial intelligence. These scenarios may be absolutely low probability but should be accounted for in nanotechnology policy given the existential nature of the threat. Beyond this, I think that Drexler's goal of addressing misconceptions about out-of-control nano-machines may also fail to recognize risks from out-of-control nano-machines which are outside the purview of comparatively limited APM. I have no reasoning for this beyond the sense that any theoretically possible existential risk ought to merit closer consideration and that many people have given closer consideration to it.

So in summary, APM is a lot more likely than I initially expected and I also suspect it could be a lot more dangerous as well. Let me know your thoughts!

Why don’t we treat geniuses like professional athletes?

I mean in the work-environment sense, rather than the celebrity-and-endorsement-deals sense.

A professional sports team gets a lot of benefits that are designed to keep physically talented people performing at their peak. They have support for things like recovery, sleep, nutrition, physical fitness, and of course the specific skills they use to play the game. This support takes the form of specialized personnel who are also employed by the team, whose job it is to work with the athletes for those purposes. The whole environment is geared towards performance maintenance.

The basics of sleep, diet and exercise are required for optimal performance across all domains; being in an environment that optimizes for them is an advantage.

Sports teams balance practice and games. Practice still makes sense for knowledge work; games analogize to projects/development/etc.

Decomposing practices is where the interesting bits might be. In athletics, there is almost always a ready-made set of drills for any particular skillset, which can be prioritized by a coach working with an athlete. There is already a notion of managing the workload: athletes have overtraining and thinkers have burnout.

In The Power of the Context Alan Kay describes funding people over projects and orienting them with a vision in lieu of specific goals. This sounds suspiciously like persuading geniuses to act as though they were on a team. Alan makes the comparison himself:

> “Our game is more like art and sports than accounting, in that high percentages of failure are quite OK as long as enough larger processes succeed."

I am tempted to go further and say that in the context of things like science or math research failures are still a positive contribution insofar as they establish that something doesn’t work, which makes future attempts more likely. This isn’t much of a thing in sports, which are entirely built around repeatable object-level activities; striking out does not make the next batter more likely to hit.

It seems like the object vs meta level distinction also highlights where the analogy breaks down. In a game like football, each player has a specialized role which contributes to moving the ball down the field. They can spend time mastering a set of moves which they adapt on the fly and can be relied on consistently. There is not a clear place for that in service of the PARC vision of “interactive computing as a complementary intellectual partner for people pervasively networked world-wide”.

We could torture the analogy ruthlessly and say that visions are meta-goals and that a meta-athlete could practice their meta-skills of finding how to complement the team’s work in advancing the vision. That even sounds sort of plausible, until you come up against the question of how to define those meta-skills. I feel like a checklist that goes like "Can you solve a simplified version of the problem? Can you generalize a solution from a similar problem?" doesn't seem to cut it. Though such a checklist would hardly be a bad idea. Based on The Rocket Alignment Problem I envision motivational posters

hanging around that go "The beacons are lit! Gondor calls for A.I.D: **A**rticulate the confusion; **I**solate the confusion; **D**issolve the confusion." Also, I am reminded of a talk given by Gian-Carlo Rota called Ten Lessons I Wish I Had Been Taught, in which he makes note of the fact that Erdos and Hilbert both employed a few tricks consistently in most of their work. Perhaps the role of the head meta-coach is to choose people such that their tricks complement each other well.

So we have the problem of identifying what skills exactly should be the focus of training and practice. We also have the problem of a shortage of personnel to serve as trainers and coaches, though I wonder if this is as intractable as it seems at first blush. The lion's share of training is not so much mastery of the skill yourself as having an outside-and-informed perspective on someone else's execution. The question is, how much is enough? A second question is, could this be automated instead?

Deconstructing Biases In Media

This is a linkpost for http://umich.edu/~newsbias/index.html

I found this website via the United Nations Educational Scientific and Cultural Organization's media literacy instagram which very nicely goes through contrasting news articles about the same event and picks apart the differences and what they imply. Unlike other sites, this one is very straightforward and clear about which bias they are tackling and how it's shown in their examples. It also has some good activities about bias.

New /r/gwern subreddit for link-sharing

This is a linkpost for https://www.reddit.com/r/gwern/

On insecurity as a friend

There’s a common narrative about confidence that says that confidence is good, insecurity is bad. It’s better to develop your confidence than to be insecure. There’s an obvious truth to this.

But what that narrative does not acknowledge, and what both a person struggling with insecurity and their well-meaning friends might miss, is that that insecurity may be in place for a reason.

You might not notice it online, but I’ve usually been pretty timid and insecure in real life. But this wasn’t *always* the case. There were occasions earlier in my life when I was less insecure, more confident in myself.

I was also pretty horrible at things like reading social nuance and figuring out when and why someone might be offended. So I was given, repeatedly, the feedback that my behavior was bad and inappropriate.

Eventually a part of me internalized that as “I’m very likely to accidentally offend the people around me, so I should be very cautious about what I say, ideally saying nothing at all”.

This was, I think, the *correct* lesson to internalize at that point! It shifted me more into an observer mode, allowing me to just watch social situations and learn more about their dynamics that way. I still don’t think that I’m *great* at reading social nuance, but I’m at least better at it than I used to be.

And there have been times since then when I’ve decided that I should act with more confidence, and just get rid of the part that generates the insecurity. I’ve been about to do something, felt a sense of insecurity, and walked over the feeling and done the thing anyway.

Sometimes this has had good results. But often it has also led to things blowing up in my face, with me inadvertently hurting someone and leaving me feeling guilty for months afterwards.

Turns out, that feeling of insecurity wasn’t a purely bad thing. It was throwing up important alarms which I chose to ignore, alarms which were sounding because it recognized my behavior as matching previous behavior which had had poor consequences.

Yes, on many occasions that part of me makes me way *too* cautious. And it would be good to moderate that caution a little. But the same part which generates the feelings of insecurity is the same part which is constantly working to model other people and their experience, their reactions to me. The part that is doing its hardest to make other people feel safe and comfortable around me, to avoid doing things that would make them feel needlessly hurt or upset or unsafe, and to actively let them know that I’m doing this.

Just carving out that part would be a mistake. A moral wrong, even.

The answer is not to get rid of it. The answer is to integrate its cautions better, to keep it with me as a trusted friend and ally – one which feels safe enough about

getting its warnings listened to, that it will not scream all the time just to be heard.

Kalman Filter for Bayesians

Summary: the Kalman Filter is Bayesian updating applied to systems that are changing over time, assuming all our distributions are Gaussians and all our transformations are linear.

Preamble - the general Bayesian approach to estimation: the Kalman filter is an approach to estimating moving quantities. When I think about a Bayesian approach to estimation, I think about passing around probability distributions: we have some distribution as our prior, we gather some evidence, and we have a new distribution as our posterior. In general, the mean of our distribution measures our best guess of the underlying value, and the variance represents our uncertainty.

In the Kalman filter, the only distribution we use is the normal/Gaussian distribution. One important property of this is that it can be parameterized completely by the mean and variance (or covariance in the multi-variate case.) If you know those two values, you know everything about the distribution.

As a result, people often talk about the Kalman filter as though it's estimating means and variances at different points, but I find it easier to think of it as outputting a distribution representing our current knowledge at any point.

The simplest case: taking multiple measurements of a fixed quantity with an accurate but imprecise sensor. For example, say we're trying to measure the temperature with a thermometer that we believe is accurate but has a variance of 5 degrees^2.

We're very bad at estimating temperatures by hand, so let's say our prior distribution is that the temperature is somewhere around 70 degrees with a variance of 20, or $N(70, 20)$. We take one readout from the thermometer, which (by assumption) yields a normal distribution centered around the true temperature with variance 5: N(t, 5). The thermometer reads 78. What's our new estimate?

Well, it turns out there's a simple rule for combining Normal distributions with known variance: if our prior is $N(\mu_0, \sigma_0^2)$ and our observation is $N(\mu_1, \sigma_1^2)$ then the posterior has mean

(1) $\mu' = \mu_0 + k(\mu_1 - \mu_0)$

(2) $\sigma'^2 = \sigma_0^2 - k\sigma_0^2$, where

(3) $k = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}$ is called the Kalman gain.

So if our first reading is 72, then k is $\frac{20}{25} = .8$, $\sigma'^2 = 20 - .8 * 20 = 4$, and $\mu' = 70 + .8 * (78 - 70) = 76.4$. If we take another reading, we'd apply the same set of calculations, except our prior would be $N(76.4, 4)$.

Some intuition: let's look at the Kalman gain. First, note that its value is always between 0 or 1. Second, note that the gain is close to 0 if σ_1^2 is large compared to σ_0^2 , and close to 1 in the opposite case. Intuitively, we can think of the Kalman gain as a ratio of how much we trust our new observation relative to our prior, where the variances are a measure of uncertainty.

What happens to the mean? It moves along the line from our prior mean to the observation. If we trust the observation a lot, k is nearly 1, and we move almost all the way. If we trust the prior much more than the observation, we adjust our estimate very little. And if we trust them equally, we take the average of the two.

Also note that the variance always goes down. Once again, if we trust the new information a lot, the variance goes down a bunch. If we trust the new information and our prior equally, then the variance is halved.

Finally, as a last tidbit, it doesn't matter whether which distribution is the prior and which is the observation in this case - we'll get exactly the same posterior if we switch them around.

Adding a sensor: none of the math above assumes we're always using the same sensor. As long as we assume all our sensors draw from distributions centered around the true mean and with a known (or estimated) variance, we can update on observations from any number of sensors, using the same update rule.

Measuring multiple quantities: what if we want to measure two or more quantities, such as temperature and humidity? Then we now have multivariate normal distributions. While a single-variable Gaussian is parameterized by its mean and variance, an n-variable Gaussian is parameterized by a vector of n means and an $n \times n$ covariance matrix: $N(\vec{\mu}, \Sigma)$.

Our update equations are the multivariate versions of the equations above: given a prior distribution $N(\vec{\mu_0}, \Sigma_0)$ and a measurement $\vec{\mu_1}$ from a sensor with covariance matrix Σ_1, our posterior distribution is $N(\vec{\mu'}, \Sigma')$ with:

(4) $\vec{\mu'} = \vec{\mu_0} + K\vec{\mu_1}$

(5) $\Sigma' = \Sigma_0 - K\Sigma_0$

(6) $K = \Sigma_0(\Sigma_0 + \Sigma_1)^{-1}$

These are basically just the matrix versions of equations (1), (2), and (3).

Adding predictable change over time: so far, we've covered Bayesian updates when you're making multiple measurements of some static set of quantities. But what about when things are changing? A classic example is a moving car. For this case, let's assume we're measuring two quantities – position and velocity.

For a bit more detail, say at time 0 our vector $\vec{\mu_0} = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}$ where x_0 is the position and v_0 is velocity. Then at time τ, we might expect the position to be $x_0 + \tau \cdot v_0$, and the velocity to be the same on average. We can represent this with a matrix: $\vec{\mu'} = F\vec{\mu_0}$, where F is the matrix $\begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix}$.

More generally, say our belief at time t is $N(\vec{\mu_0}, \Sigma_0)$. Then our belief at time $t + \tau$, before we make any new observations, should be $FN(\vec{\mu_0}, \Sigma_0)$. Fortunately there's a simple formula for this: $FN(\vec{\mu_0}, \Sigma_0) = N(F\vec{\mu_0}, F\Sigma_0 F^T)$.

Putting it all together, say our belief at time t is $N(\vec{\mu_0}, \Sigma_0)$, and at time $t + \tau$ we measure a value $\vec{\mu_1}$ from a sensor with covariance matrix Σ_1, then we perform the Bayesian update with $FN(\vec{\mu_0}, \Sigma_0) = N(F\vec{\mu_0}, F\Sigma_0 F^T)$ as the prior and $N(\vec{\mu_1}, \Sigma_1)$ as the posterior:

(7) $\vec{\mu'} = F\vec{\mu_0} + K\vec{\mu_1}$

(8) $\Sigma' = F\Sigma_0 F^T - K\Sigma_0$

(9) $K = F\Sigma_0 F^T (F\Sigma_0 F^T + \Sigma_1)^{-1}$

And that's the main idea! We just adjust our prior by applying a transition function/matrix to it first In practice, the Kalman filter tends to quickly converge to true values, and is widely used in applications such as GPS tracking.

Thoughts on short timelines

This is a linkpost for http://s-risks.org/thoughts-on-short-timelines/

Coordination Problems in Evolution: The Rise of Eukaryotes

Introduction

This is a series of posts about coordination problems, as they appear in the course of biological evolution. It is based on the book "The Major Transitions in Evolution" by John Maynard Smith and Eörs Szathmáry. Previous part, discussing Eigen's paradox as well as the origin of chromosomes, can be found here.

In this part we are going to look at the origin of eukaryotic cell, specifically at its acquisition of endosymbiotic organelles, and at the origin of multicellularity.

Prokaryota vs. eukryota

While all single-celled organisms may look like similar wiggly little creatures to us, there is a huge difference between prokaryota (like bacteria) and eukaryota (like protozoa or, for that matter, our own cells). The cell wall is different. The interior of the cell is different. One has rigid cell wall, in the other it's the cytoskeleton that holds the cell together. One has a single-origin DNA strand attached to the cell wall, the other has nucleus containing chromosomes. One has mitochondria and chloroplasts, the other does not. Even the mechanism of cell division is different. If we haven't known that we share part of the genome, it would be easy to make a mistake and believe that the life on Earth had originated at two separate occasions.

The transition from prokayotes to eukaryotes is likely the most complex transition in the entire course of evolution. It took two billion years to happen. More than the emergence of life itself.

All that being said, we are going to look only at a single part of the evolution of eukaryotes, namely at their acquisition of mitochondria. Mitochondria were free-living cells once. But then they've became an inseparable part of eukaryotic cell. Hence, back to the coordination problems!

How did it come to be that some cells started living within other cells? Well, assuming that flexible cell wall and phagocytosis evolved before the domestication of mitochondria, getting them inside wouldn't be a big problem. It happens each time one cell eats another.

What's more interesting is how did the guest cell survive and how did the cooperative behavior between the host and the guest evolve.

Symbiosis

Let's make a digression and think about symbiosis for a second. If we assume that there are only two strategies for the host (cultivate the symbiont or try to kill it) and two strategies for the symbiont (cooperate with the host or parasitize) then the problem gets reduced to variants of the prisoner's dilemma game.

Consider this kind of arrangement of payoffs. The numbers specify the fitness of the host (left) and the symbiont (right):

Host / Symbiont	cooperate	parasitize
cultivate	20 / 20	5 / 30
kill	15 / 0	**10 / 5**

It can be easily seen that there is only one equilibrium: Whatever the host does it's better for the symbiont to parasitize. And if the symbiont is a parasite it's always better for the host to kill it.

Under what conditions do we see this kind of game? The authors point out that this happens when **each individual host acquires some genetically different symbionts from the environment**. The reason is that it doesn't pay for the symbiont to invest in the cooperation with the host if the host is going to be killed by a different symbiont anyway.

How about a different scenario?

Host / Symbiont	cooperate	parasitize
cultivate	**20 / 20**	5 / 10
kill	15 / 0	**10 / 5**

The ideal strategy for the symbiont is not clear in this case. If the host is cultivator it may pay to the symbiont to cooperate. If, on the other hand, the host tries to kill it, the best thing for the symbiont to do would be to multiply as fast as possible, regardless to any damage to the host.

This kind of setup is expected **if each host acquires only a single symbiont from the environment**:

> However, with hosts infected by a single symbiont, cooperative mutualism is likely to be stable once it evolves. The evolution from parasitism to mutualism will be favored if the hosts killing response is ineffective, and if the further spread of the symbiont is greater if the host does survive. It will not occur if the host can rapidly rid itself of the parasite, or if the parasite spreads only by killing the host.

Finally, let's have a look at the the following scenario:

Host / Symbiont	cooperate	parasitize
cultivate	**20 / 20**	5 / 10
kill	15 / 15	10 / 0

Again, there's only one equilibrium. It's always better for the symbiont to cooperate and once it's cooperating, it's better for the host to cultivate it.

This happens when **the host acquires one or a few symbionts from one of its parents**.

It makes sense: If the only place you can disperse to are your host's children you really don't want to kill it.

So there's a rule of thumb emerging here: If the transmission of the symbiont happens between unrelated individuals (horizontal transmission) the symbiosis will evolve towards parasitism. If the symbiont is passed only from one parent to its children, then the relationship will evolve towards mutualism.

In fact, both experiments and observations in the wild show that vertical transmission of the symbiont leads to mutualism and horizontal transmission leads to parasitsm. There are some exceptions though. For example, the transmission of luminous bacteria in deep-see fish is horizontal, yet the symbionts are essential to the survival of their hosts.

Parasites or livestock?

Now, let's get back to the origin of eukaryotic cell. What was the relationship between the early host cells and early mitochondria?

It may have been that the mitochondria were parasites. Maybe they sometimes escaped the host cell and infected different cells. However, the authors hint at an interesting alternative: The host cells may have farmed the mitochondria for the later consumption, just like we do with the cattle.

One important point to understand here is that, however we feel about slaughtering cows, from the population generics point of view it's a mutually beneficial arrangement. Homo Sapiens gets steaks. Bos Taurus becomes one of the most common terrestrial vertebrates around.

So, the host cells may have first consumed mitochondria, but then learned to keep them around (or rather inside) so that they can be consumed later.

And we do see some evidence that the host cell adopts active measures to keep the relationship mutualistic. In sexually reproducing species the transmission of mitochondria happens from one parent only. When human egg merges with human sperm, all the mitochondria from the sperm are discarded and only those from the egg make their way into the embryo. That, according to the model described above, prevents competition between the different strains of mitochondria at the expense of the host cell.

Later on in the evolution, straightforward consumption of the symbionts must have been replaced by protein "taps" that we see installed into the cell wall of mitochondria today. The taps allow the nutrients produced by the mitochondria to flow into the host cell. Think of Maasai puncturing the flesh of a cow and drinking the blood without killing the animal. The fact that the tap protein is always encoded in the DNA of the host cell rather than in mitochondrial DNA is a hint that the idea of the host cells "farming" mitochondria may not be that implausible.

Gene transfer to the nucleus

Once the mitochondria were living inside the host cell a curious process began. The genes from the mitochondria started "jumping" into the host cell's nucleus.

By losing their genes, mitochondria lost the chance to break free from the eukaryotic cell for good. So why did it happen? And who benefited?

I really like this process because it shows how complex the interplay between different levels of selection can be. In particular, we have to do with three distinct levels of selection here: Selection on the level of the host cell, selection on the level of mitochondria and selection on the level of a single mitochondrial gene.

First, we can imagine a mitochondrial gene getting attached to a nuclear chromosome. It would be clearly advantageous for the gene: One more copy! Hooray! What's not to like?

But why didn't the gene got discarded from the nuclear DNA given that it performed no useful function? Well, it turns out that nuclear DNA can contain humungous amounts of dead code and yet the code doesn't get discarded by natural selection. Contrast that with prokaryotes which tend to keep their genetic code short, sweet and streamlined.

But wait. The gene would still be translated into protein. That would be a useless expenditure of energy and thus it would be selected against. To make it advantageous for the cell there must have been a mechanism to transport the protein back into mitochondria. Luckily, all that is needed for that is to add to the protein a "transit peptide", a handle which would be recognized by a receptor in the mitochondrial membrane and used to carry the protein inside. Creating such a handle is easy. Baker & Schatz pasted randomly chosen pieces of E. coli and mouse DNA in front of protein genes, and found that 2.7 per cent of the bacterial inserts and 5 per cent of the mammalian ones were successful transit peptides.

Another hint that the transition may be easy is that there is no distinct pattern to the transit peptides. In other words, the transit peptides probably evolved 700 times independently — once for each mitochondrial gene that was transferred into the nucleus.

When the gene is in nucleus and the peptide handle in its place, the mitochondria can gain advantage by discarding the genes that they don't need anymore (they are importing those proteins from the outside anyway). And, as already mentioned, prokaryotes are very good at stripping their genome of any unnecessary baggage. Nick Lane does some back-of-the-envolope calculations and concludes that the energy savings are truly huge.

All of that being the case, the question is rather why all the genes haven't been transferred to the nucleus.

Why the gene transfer stopped

For mitochondria, the process was stopped by the change in mitochondrial genetic code (see only kind-of-related but fun-to-read column by Douglas Hofstadter). As soon as one of the mitochondrial codons began coding for a different amino acid, the genes could no longer jump to the nucleus. When they did they were turned into defective proteins by the old, unmodified nuclear translation machinery.

But that can't be the entire story. The chloroplast genes are encoded in the plain old generic code. Yet the chloroplasts still keep some of the genes for themselves. This may (or may not) indicate that there's still some level of separate identity to the organelles, that they may have goals of their own not fully aligned with the goals of the enclosing eukaryotic cell. An example of that would be, for example, mitochondria trying to distort the sex ratio of the host species.

(As a side note, there are organelles called peroxisomes that were once thought to be endosymbionts, very much like mitochondria. Except that they had no genes at all. It has been suggested that they may be endosymbionts that have transferred all of their genes to the nucleus. However, that idea has been recently challenged.)

Multicullular life and Orgel's second rule

Multicellular life sure looks like it has a coordination problem. All those billions and trillions of cells have to cooperate somehow. Most of them have to give up individual reproduction and rather work for the benefit of all. Hell, there's even programmed cell death where the cell is expected to willingly die when there's no use for it any more.

But when you take a step back the argument doesn't make sense. All those cells are genetically identical. There aren't multiple entities engaging in a coordination problem. There's just one entity: The multi-cellular organism itself.

Or is there?

It may be instructive to pause here for a while and do an exercise in evolutionary thinking.

Consider what happens if a somatic cell mutates.

The mutation may cause the cell to divide in unregulated manner.

But there's even more intriguing possibility: Imagine that a the cell mutates in such a way that it's more likely to give rise to a germ cell. For example, a plant cell that would otherwise produce a leaf would give rise to a flower instead. By doing so it would lower the overall fitness of the organism: The plant would now have less leaves than what's optimal. However, at the same time the renegade cell would increase its own fitness because any pollen or seed produced by that flower would carry the mutated gene instead of the original one.

So what do you think? Does the above make sense or does it not? Is it really an intra-organism conflict? Think about that. I'll wait.

...

Smith and Szathmáry approach the problem by splitting it into two parts.

First, they discuss whether mutation that increases a chance of giving rise to a gamete creates an internal conflict and, consequently, selection pressure for other cells to evolve a mechanism to prevent such mutations. They conclude that it doesn't. After all, this is not much different from when the mutation occurs in a germ cell. The child will be slightly genetically different from the parent, but that's just how evolution works. If the child happens to be more fit than the parent, it will eventually prevail in the competition on the organism level. If not so, the new strain will be eliminated by the natural selection.

The second part of the question is what happens if the mutant is malignant, i.e. if it causes unregulated cell proliferation. We call that cancer. In that case, the authors conclude, there will be an actual selective force to prevent, or delay, the malignancy.

Have you got that right? If not so, don't be disappointed and think about Orgel's second rule. The rule says: "Evolution is cleverer than you are."

If you want to remember just one thing about evolution, Orgel's second rule may be a good choice.

In fact, Smith and Szathmáry, as good evolutionary biologists, have the rule ingrained and conclude the section by hedging their bets:

> There is, therefore, no reason to think that [specific mechanisms discussed in the book] evolved to suppress cell-cell competition. But the question is important, and we do not regard our arguments as decisive.

To be continued

In the following parts I will cover the origin of sex and of social species. In the end I am going to speculate about possible parallels between coordination problems in evolution and coordination problems in human society.

Cognitive Enhancers: Mechanisms And Tradeoffs

[Epistemic status: so, so, so speculative. I do not necessarily endorse taking any of the substances mentioned in this post.]

There's been recent interest in "smart drugs" said to enhance learning and memory. For example, from the Washington Post:

> When aficionados talk about nootropics, they usually refer to substances that have supposedly few side effects and low toxicity. Most often they mean piracetam, which Giurgea first synthesized in 1964 and which is approved for therapeutic use in dozens of countries for use in adults and the elderly. Not so in the United States, however, where officially it can be sold only for research purposes. Piracetam is well studied and is credited by its users with boosting their memory, sharpening their focus, heightening their immune system, even bettering their personalities.

Along with piracetam, a few other substances have been credited with these kinds of benefits, including some old friends:

> "To my knowledge, nicotine is the most reliable cognitive enhancer that we currently have, bizarrely," said Jennifer Rusted, professor of experimental psychology at Sussex University in Britain when we spoke. "The cognitive-enhancing effects of nicotine in a normal population are more robust than you get with any other agent. With Provigil, for instance, the evidence for cognitive benefits is nowhere near as strong as it is for nicotine."

But why should there be smart drugs? Popular metaphors speak of drugs fitting into receptors like "a key into a lock" to "flip a switch". But why should there be a locked switch in the brain to shift from THINK WORSE to THINK BETTER? Why not just always stay on the THINK BETTER side? Wouldn't we expect some kind of tradeoff?

Piracetam and nicotine have something in common: both activate the brain's acetylcholine system. So do three of the most successful Alzheimers drugs: donepezil, rivastigmine, and galantamine. What is acetylcholine and why does activating it improve memory and cognition?

Acetylcholine is many things to many people. If you're a doctor, you might use neostigmine, an acetylcholinesterase inhibitor, to treat the muscle disease myasthenia gravis. If you're a terrorist, you might use sarin nerve gas, a more dramatic acetylcholinesterase inhibitor, to bomb subways. If you're an Amazonian tribesman, you might use curare, an acetylcholine receptor antagonist, on your blowdarts. If you're a plastic surgeon, you might use Botox, an acetylcholine release preventer, to clear up wrinkles. If you're a spider, you might use latrotoxin, another acetylcholine release preventer, to kill your victims – and then be killed in turn by neonictinoid insecticides, which are acetylcholine agonists. Truly this molecule has something for everybody – though gruesomely killing things remains its comparative advantage.

But to a computational neuroscientist, acetylcholine is:

> …a neuromodulator [that] encodes changes in the precision of (certainty about) prediction errors in sensory cortical hierarchies. Each level of a processing hierarchy sends predictions to the level below, which reciprocate bottom-up signals. These signals are prediction errors that report discrepancies between top-down predictions and representations at each level. This recurrent message passing continues until prediction errors are minimised throughout the hierarchy. The ensuing Bayes optimal perception rests on optimising precision at each level of the hierarchy that is commensurate with the environmental statistics they represent. Put simply, to infer the causes of sensory input, the brain has to recognise when sensory information is noisy or uncertain and down-weight it suitably in relation to top-down predictions.

…is it too late to opt for the gruesome death? It is? Fine. God help us, let's try to understand Friston again.

In the predictive coding model, perception (maybe also everything else?) is a balance between top-down processes that determine what you should be expecting to see, and bottom-up processes that determine what you're actually seeing. This is faster than just determining what you're actually seeing without reference to top-down processes, because sensation is noisy and if you don't have some boxes to categorize things in then it takes forever to figure out what's actually going on. In this model, acetylcholine is a neuromodulator that indicates increased sensory precision – ie a bias towards expecting sensation to be signal rather than noise – ie a bias towards trusting bottom-up evidence rather than top-down expectations.

In the study linked above, Friston and collaborators connect their experimental subjects to EEG monitors and ask them to listen to music. The "music" is the same note repeated again and again at regular intervals in a perfectly predictable way. Then at some random point, it unexpectedly shifts to a different note. The point is to get their top-down systems confidently predicting a certain stimulus (the original note repeated again for the umpteenth time) and then surprise them with a different stimulus, and measure the EEG readings to see how their brain reacts. Then they do this again and again to see how the subjects eventually learn. Half the subjects have just taken galantamine, a drug that increases acetylcholine levels; the other half get placebo.

I don't understand a lot of the figures in this paper, but I think I understand this one. It's saying that on the first surprising note, placebo subjects' brains got a bit more electrical activity [than on the predictable notes], but galantamine subjects' brains got much more electrical activity. This fits the prediction of the theory. The placebo subjects have low sensory precision – they're in their usual state of ambivalence about whether sensation is signal or noise. Hearing an unexpected stimulus is a little bit surprising, but not completely surprising – it might just be a mistake, or it might just not matter. The galantamine subjects' brains are on alert to expect sensation to be very accurate and very important. When they hear the surprising note, their brains are very surprised and immediately reevaluate the whole paradigm.

One might expect that the very high activity on the first discordant note would be matched with lower activity on subsequent notes; the brain has now fully internalized the new prediction (ie is expecting the new note) and can't be surprised by it anymore. As best I can tell, this study doesn't really show that. A very similar study by some of the same researchers does. In this one, subjects on either galantamine or a

placebo have to look at a dot as quickly as possible after it appears. There are some arrows that might or might not point in the direction where the dot will appear; over the course of the experiment, the accuracy of these arrows changes. The researchers measured how quickly, when the meaning of the arrows changed, the subjects shifted from the old paradigm to the new paradigm. Galantamine enhanced the speed of this change a little, though it was all very noisy. Lower-weight subjects had a more dramatic change, suggesting an effective dose-dependent response (ie the more you weigh, the less effect a constant-weight dose of galantamine will have on your body). They conclude:

> This interpretation of cholinergic action in the brain is also in accord with the assumption of previous theoretical notions posing that ACh controls the speed of the memory update (i.e., the learning rate)

“Learning rate” is a technical term often used in machine learning, and I got a friend who is studying the field to explain it to me (all mistakes here are mine, not hers). Suppose that you have a neural net trying to classify cats vs. dogs. It’s already pretty well-trained, but it still makes some mistakes. Maybe it’s never seen a Chihuahua before and doesn’t know dogs can get that small, so it thinks “cat”. A good neural network will learn from that mistake, but the amount it learns will depend on a parameter called learning rate:

If learning rate is 0, it will learn nothing. The weights won’t change, and the next time it sees a Chihuahua it will make the exact same mistake.

If learning rate is very high, it will overfit. It will change everything to maximize the likelihood of getting that one picture of a Chihuahua right the next time, even if this requires erasing everything it has learned before, or dropping all “common sense” notions of dog and cat. It is now a “that one picture of a Chihuahua vs. everything else” classifier.

If learning rate is a little on the low side, the model will be very slow to learn, though it will eventually converge on a good understanding of its topic.

If learning rate is a little on the high side, the model will learn very quickly, but “jump around” between different understandings heavily weighted toward what best fits the last case it has worked on.

On many problems, it’s a good idea to start with a high learning rate in order to get a basic idea what’s going on first, then gradually lower it so you can make smaller jumps through the area near the right answer without overshooting.

Learning rates are sort of like sensory precision and bottom-up vs. top-down weights, in that as a high learning rate says to discount prior probabilities and weight the evidence of the current case more strongly. A higher learning rate would be appropriate in a signal-rich environment, and a lower rate appropriate in a noise-rich environment.

If acetylcholine helps set the learning rate in the brain, would it make sense that cholinergic substances are cognitive enhancers / “study drugs”?

You would need to cross a sort of metaphorical divide between a very mechanical and simple sense of “learning” and the kind of learning you do where you study for your final exam on US History. What would it mean to be telling your brain that your US

History textbook is “a signal-rich environment” or that it should be weighting its bottom-up evidence of what the textbook says higher than its top-down priors?

Going way beyond the research into total speculation, we could imagine the brain having some high-level intuitive holistic sense of US history. Each new piece of data you receive could either be accepted as a relevant change to that, or rejected as “noise” in the sense of not worth updating upon. If you hear that the Battle of Cedar Creek took place on October 19, 1864 and was a significant event in the Shenandoah Valley theater of the Civil War, then – if you’re like most people – it will have no impact at all on anything beyond (maybe, if you’re lucky) being able to parrot back that exact statement. If you learn that the battle took place in 2011 and was part of a Finnish invasion of the US, that changes a lot and is pretty surprising and would radically alter your holistic intuitive sense of what history is like.

Thinking of it this way, I can imagine these study drugs helping the exact date of the Battle of Cedar Creek seem a little bit more like signal, and so have it make a little bit more of a dent in your understanding of history. I’m still not sure how significant this is, because the exact date of the battle isn’t surprising to me in any way, and I don’t know what I would update based on hearing it. But then again, these drugs have really subtle effects, so maybe not being able to give a great account of how they work is natural.

And what about the tradeoff? Is there one?

One possibility is no. The idea of signal-rich vs. signal-poor environments is useful if you’re trying to distinguish whether a certain pattern of blotches is a camoflauged tiger. Maybe it’s not so useful for studying US History. Thinking of Civil War factoids as anything less than maximally-signal-bearing might just be a mismatch of evolution to the modern environment, the same way as liking sweets more than vegetables.

Another possibility is that if you take study drugs in order to learn the date of the Battle of Cedar Creek, you are subtly altering your holistic intuitive knowledge of American history in a disruptive way. You’re shifting everything a little bit more towards a paradigm where the Battle of Cedar Creek was unusually important. Maybe the people who took piracetam to help them study ten years ago are the same people who go around now talking about how the Civil War explains every part of modern American politics, and the 2016 election was just the Confederacy getting revenge on the Union, and how the latest budget bill is just a replay of the Missouri Compromise.

And another possibility is that you’re learning things in a rote, robotic way. You can faithfully recite that the Battle of Cedar Creek took place on October 19, 1864, but you’re less good at getting it to hang together with anything else into a coherent picture of what the Civil War was really *like*. I’m not sure if this makes sense in the context of the learning rate metaphor we’re using, but it fits the anecdotal reports of some of the people who use Adderall – which has some cholinergic effects in addition to its more traditional catecholaminergic ones.

Or it might be weirder than this. Remember the aberrant salience model of psychosis, and schizophrenic Peter Chadwick talking about how one day he saw the street “New King Road” and decided that it meant Armageddon was approaching, since Jesus was the new king coming to replace the old powers of the earth? Is it too much of a stretch to say this is what happens when your learning rate is much too high, kind of like the neural network that changes everything to explain one photo of a Chihuahua? Is this why nicotine has weird effects on schizophrenia? Maybe higher learning rates can

promote psychotic thinking – not necessarily dramatically, just make things get a little weird.

Having ventured this far into Speculation-Land, let's retreat a little. Noradrenergic and Cholinergic Moduation of Belief Updating does some more studies and fails to find any evidence that scopolamine, a cholinergic drug, alters learning rate (but why would they use scopolamine, which acts on muscarinic acetylcholine receptors, when every drug suspected to improve memory act on nicotinic ones?). Also, nicotine seems to help schizophrenia, not worsen it, which is the opposite of the last point above. Also, everything above about acetylcholine sounds kind of like my impression of where dopamine fits in this model, especially in terms of it standing for the precision of incoming data. This suggests I don't understand the model well enough for everything not to just blend together to me. All that my usual sources will tell me is that the acetylcholine system modulates the dopamine system.

(remember, neuroscience madlibs is "_____ modulates _____", and no matter what words you use the sentence is always true.)

Epistemic Spot Check: The Dorito Effect (Mark Schatzker)

Epistemic Spot Checks is a series in which I fact check claims a book makes, to determine its trustworthiness. It is not a book review or a check on every claim the book makes, merely a spot check of what I find particularly interesting or important (or already know).

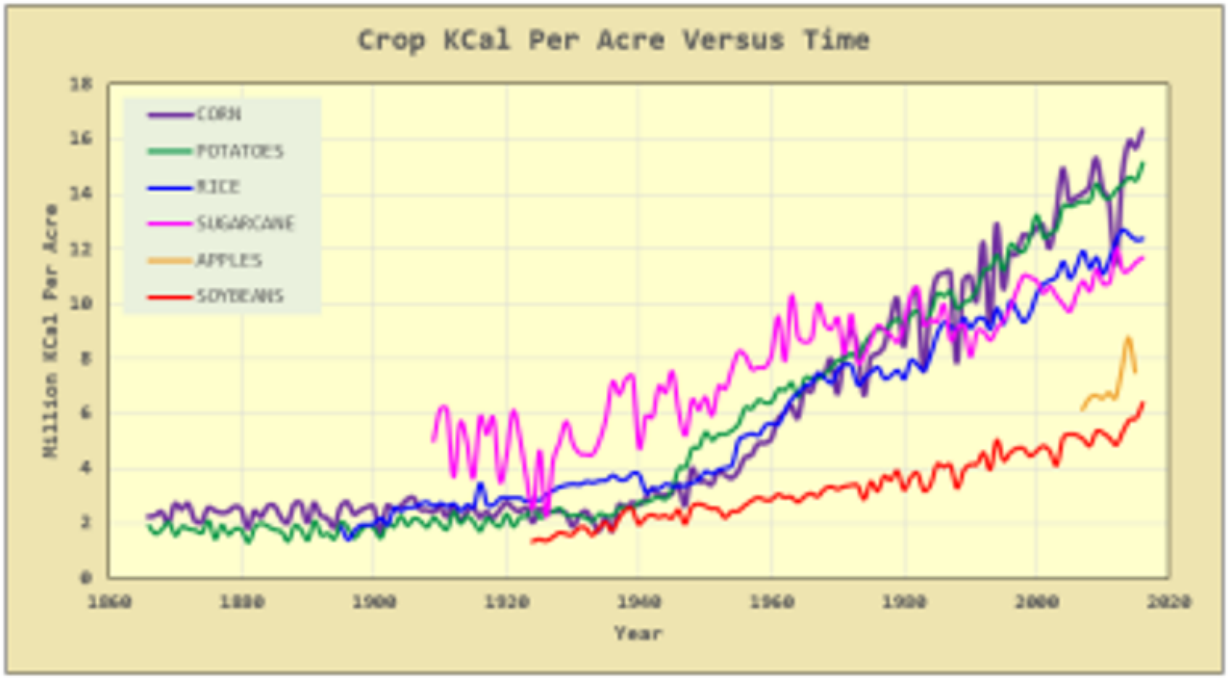
Today's subject is *The Dorito Effect*, which claims that Americans are getting fat because food is simultaneously getting blander and less nutritious, and then more intensely flavored through artificial means. This is leaving people fat and yet malnourished.

Claims

Claim: Humans did not get fatter over the last 100 years due to changes in genetics.
True. People are fatter than their ancestors, indicating it's not a change in genetics (although genetics still plays a role in an individual's weight).

Claim: Casimir Funk discovered that an extract of brown rice could cure beriberi in chickens.
True.

Claim: In 1932, the average farm produced 63 sacks of potatoes/acre. By the mid 1960s, it was 200 sacks/acre.
True.



(source).

Claim: Everything is getting blander and more seasoned.
More seasoned.
Blander food.
Note that both sources were provided by the book itself.

Claim: "*We eat for one reason: because we love the way food tastes. Flavor is the original craving".*
This doesn't jive with my personal experience. I definitely crave nutrients and am satisfied by them even without tasting them.

Claim: "In 1946 and 1947, regional Chicken Of Tomorrow contests were held."
True.

Claim: *Over time the Chicken Of Tomorrow winners consistently weighed more, with less feed and less time to maturity.*
True.

Claim: Produce is getting less nutritious over time.
True (source provided by author).

Conclusions

Extremely trustworthy, and therefore worrisome, given the implication that food is becoming inexorably worse. *Dorito Effect* is unfortunately light on solutions, so you might just freak yourself out to no purpose. On the other hand, if you're looking for a kick to start eating better, this could easily be it.

Decision Theory Anti-realism

This is crossposed from my blog. While I think the ideas here are solid I think the presentation still needs some work so I'd also appreciate comments on the presentation so I can turn this into a more polished essay, e.g., is the second section worth keeping and what should be expanded/cut.

With the recent flurry of posts in the rationalist community about which decision theory (e.g. CDT EDT UDT etc..) it's time to revisit the theme of this blog: rejecting rationality realism In this case that means pointing out that there isn't actually a well-defined fact of the matter about which decision theory is better. Of course, nothing stops us from arguing with each other about the best decision theory but those disagreements are more like debates about what's the best programming language than disagreements about the chemical structure of Benzene.. , ,

Any attempt to compare decision theories must first address the question: **what does it mean for one decision theory to be better than another?** Unlike many pseudo-problems there is a seemingly meaningful answer to this question: **one decision theory is better than another to the extent that the choices it recommends lead to better outcomes for the agent.** Other than some ambiguity about which theory is better if neither dominates the other it seems like this gives a straightforward criteria for superiority: we just look at actual outcomes and see which decision theory offers the best results for an agent. However, this only appears to give a well-defined criteria because in every day life the subtle differences between the various ways to understand a choice and how to conceptualize making a choice don't matter.

In particular, I argue here that the kind of scenarios which distinguish between theories like EDT and CDT are hazardous places to deploy our raw intuitions and support different theories depending on how exactly one precisifies the question being asked. As such, there isn't a well-defined fact of the matter as to which decision theory is right or the best as we only *appear* to have specified a clear conception of what it means for one theory to be better than another. This doesn't mean the various decision theories aren't interesting notions to play with but it does mean there is no deep fact to get at about which one is right.

Intuitions and Motivation

So why would one even suspect there is a fact of the matter about which decision theory is best/true? I mean we wouldn't suspect there is a fact of the matter about which formal model of computation (Turing machines, register machines, DFAs, Quantum Turing Machines etc..) is the true model. They are all valid formal constructs and we use whatever one seems most appropriate to a given question. So why would we suspect decision theory to be any different?

My best guess is that people have some intuition like the following:

> I can mentally walk through the predictions of various decision theories (even in Newcomb style problems) and seeing who gets what makes it seem obvious that

one theory is better than another. Nothing about what counts as a choice or how I idealize those choices seems to make a difference.

To convince you that's not enough let me provide an extreme dramatization of how our intuitive conception of what choice would be better for you or I to make can come apart from the apparent formal payouts of a decision theory. Consider the following Newtonian, rather than Newcombian, problem. *You fall off the top of the empire State building what you do as you fall past the fifth floor?* What would one say about the virtues of Floating Decision Theory which (otherwise being equal to, say, CDT) tells us that in such a situation we should make the choice to float gently to the ground. Now obviously, one would prefer to float rather than fly but posing the problem as a decision between these two choices doesn't render it a real choice. Obviously, there is something dubious about evaluating your decision theory based on its performance on the float/fall question.

The point of this example is to illustrate that when the supposed choices given by our idealization are sufficiently different from what we normally think of as a choice we can't punt on preciscifying what it means for a choice (or a decision theory) to be better than another and simply refer back to our informal intuitions about choices. After all choosing to float doesn't make sense as a choice in our intuitive sense so how can we assume that intuitive sense steps in and tells us what it means for one theory to be better than another in such a situation?

Yet, this is precisely the kind of situation we encounter in the original Newcomb problem as the very assumption of predictability which allows the demon (or in Yudkowsky's formulation Omega) to favor the 1 boxers ensures the physical impossibility of choosing any number of boxes other than what you did choose. Of course, the same is (up to quantum mechanical randomness) true of any actual `choice' by a real person but under certain circumstances we find it useful to *idealize* it as free choice. What's different about the Newcomb problem is that, understood naively, it *simultaneously* asks us to idealize selecting 1 or 2 boxes as a free choice while assuming it isn't actually. In other words here too we can't simply punt back to our intuitions about one choice being better than another and assume it gives a clear answer.

Once we are willing to grant that it's not enough to just presume that our gut instincts about choices give rise to a well-defined notion of what it means for one decision theory to be better than another and start trying to precisify what that means it quickly becomes clear there is more than one way to do that.

Possible Precisifications

Ultimately, there is something a bit weird about asking what decision a real physical agent should take in a given situation. After all, the agent will act just as it's software dictates and/or the laws of physics require. Thus, as Yudkowsky recognizes, any comparison of decision theories is asking some kind of counterfactual. However, which counterfactual we ask makes a huge difference in what decision theory is preferable. For instance, all of the following are potential ways to precifisify the question of what it means for it to be better for XDT to be a better decision theory than YDT.

1. If there was a miracle that overrode the agent's programming/physical laws at the moment of a choice then doing so in the manner prescribed by XDT yields better

outcomes than doing so in a manner prescribed by YDT.

2. In fact those actual agents who more often choose the outcome favored by XDT do better than those who choose the outcome favored by YDT.

3. Those actual agents which adopt/apply XDT do better than those who adopt/apply YDT.

4. Suppose there is a miracle that overrode physical laws at the moment the agent's programming/internal makeup is specified then if the miracle results in outcomes more consistent with XDT than YDT the agent does better.

5. As above except with applying XDT/YDT instead of just favoring outcomes which tend to agree with it.

6. Moving one level up we could ask about which performs better, agents whose programming inclines them to adopt XDT or YDT when considered.

7. Finally, if what we are interested in is actually coding agents, i.e., writing AI software, we might ask whether programmers who code their agents to reason in a manner that prefers choice A produce agents that do better than programmers who code agents to reason in a manner that prefers choice B.

8. Pushing that one level up we could ask about whether programmers who are inclined to adopt/apply XDT/YDT as true produce agents which do better.

One could continue and list far more possibilities but these eight are enough to illustrate the point that there are multiple different kinds of questions one might want a decisions theory to answers. Importantly, depending on which one we choose we get different answers as to which theory is preferable.

For instance, note that if we are asking question 1 CDT outperforms EDT. For the purposes of question 1 the right answer to the Newcomb problem is to be a 2 boxer. After all, if we idealize the choice as a miracle that allows deviation from physical law then the demon's prediction of whether we would be a two-boxer or one-boxer no longer must be accurate so two-boxes always outperforms one boxing. It doesn't matter that your software says you will choose only one box if we are asking about outcomes where a miracle occurs and overrides that software.

On the other hand it's clearly true that EDT does better than CDT with respect to question 2. That's essentially the definition of EDT.

To distinguish the remaining options we need to consider a range of different scenarios such as demons who punish agents who actually apply/adopt XDT/YDT in reaching their conclusions. Or consider Newcombian demons who punish agents who adopt (or whose programmers adopted one of XDT/YDT). But the important point is that depending on the assumptions we make about what it means for one theory to be better than another and the kind of problems the agent will face yield different answers for the 'right' way for the agent to behave.

Ultimately, **which criteria we should use to compare decision theories depends on what we want to achieve.** Different idealizations/criteria will be appropriate depending on whether we are asking which rule we ourselves should adopt, how we should program agents to act, how we should program agents who program agents etc.. etc... At anytime we've precisified what it is we want out of our

decision theory sufficiently well to make the question of which one is the best well-defined there won't be anything left to debate about, e.g., it's analytic that CDT is the theory which yields the best outcomes if we take the nature of the agent (source code/physical makeup) to be fixed but idealize decisions as miracles that temporarily suspend the normal casual rules (allowing agents to choose things Newcombian demons predict they wouldn't).

Is Realism Impossible?

I don't take myself to have offered a proof that it's impossible to ever believe there are facts of the matter about the true decision theory. Merely offered a strong prima facie case that there probably isn't such a fact. After all, it's always possible that, like the argument that justified true belief isn't knowledge, someone will pop up and show that there really was a precise criteria for preferability/truth of decision theories implicit in the usage of every competent English speaker. But even if such a hidden fact about our linguistic commitments was found it wouldn't really tell us anything significant about the world. We would do better simply spelling out in detail what it is we seek to claim (e.g. the decisions which maximize outcome under the assumption that choices can be idealized as being little miracles) and tossing aside as, probably meaningless and at best unimportant, the question of what the best decision theory is. If there is some aspect of reality that such an attitude seems to be ignoring the burden is on those who believe this to offer evidence.