# Best of LessWrong: July 2019

1. Jeff Hawkins on neuromorphic AGI within 20 years
2. The Costs of Reliability
3. Integrity and accountability are core parts of rationality
4. A Key Power of the President is to Coordinate the Execution of Existing Concrete Plans
5. Forum participation as a research strategy
6. The Real Rules Have No Exceptions
7. No nonsense version of the "racial algorithm bias"
8. Appeal to Consequence, Value Tensions, And Robust Organizations
9. Everybody Knows
10. Does it become easier, or harder, for the world to coordinate around not building AGI as time goes on?
11. Learning biases and rewards simultaneously
12. The AI Timelines Scam
13. Magic is Dead, Give me Attention
14. What are we predicting for Neuralink event?
15. How much background technical knowledge do LW readers have?
16. Book review: The Technology Trap
17. Wolf's Dice
18. What does Optimization Mean, Again? (Optimizing and Goodhart Effects - Clarifying Thoughts, Part 2)
19. June 2019 gwern.net newsletter
20. Why artificial optimism?
21. Doublecrux is for Building Products
22. What would be the signs of AI manhattan projects starting? Should a website be made watching for these signs?
23. Self-experiment Protocol: Effect of Chocolate on Sleep
24. The Right Way of Formulating a Problem?
25. Self-consciousness wants to make everything about itself
26. Largest open collection quotes about AI
27. Black hole narratives
28. Prereq: Cognitive Fusion
29. On the purposes of decision theory research
30. Job description for an independent AI alignment researcher
31. Diversify Your Friendship Portfolio
32. Should I wear wrist-weights while playing Beat Saber?
33. Dialogue on Appeals to Consequences
34. Why did we wait so long for the bicycle?
35. How often are new ideas discovered in old papers?
36. Let's Read: Superhuman AI for multiplayer poker
37. How to take smart notes (Ahrens, 2017)
38. Are we certain that gpt-2 and similar algorithms are not self-aware?
39. What are good resources for learning functional programming?
40. Intellectual Dark Matter
41. Where is the Meaning?
42. Is there neuroscience research on cognitive biases?
43. Prediction as coordination
44. Commentary On "The Abolition of Man"
45. Do bond yield curve inversions really indicate there is likely to be a recession?
46. An Increasingly Manipulative Newsfeed
47. If I knew how to make an omohundru optimizer, would I be able to do anything good with that knowledge?
48. Robust Agency for People and Organizations
49. Do you fear the rock or the hard place?
50. How to make a giant whiteboard for $14 (plus nails)

# Best of LessWrong: July 2019

1. Jeff Hawkins on neuromorphic AGI within 20 years
2. The Costs of Reliability
3. Integrity and accountability are core parts of rationality
4. A Key Power of the President is to Coordinate the Execution of Existing Concrete Plans
5. Forum participation as a research strategy
6. The Real Rules Have No Exceptions
7. No nonsense version of the "racial algorithm bias"
8. Appeal to Consequence, Value Tensions, And Robust Organizations
9. Everybody Knows
10. Does it become easier, or harder, for the world to coordinate around not building AGI as time goes on?
11. Learning biases and rewards simultaneously
12. The AI Timelines Scam
13. Magic is Dead, Give me Attention
14. What are we predicting for Neuralink event?
15. How much background technical knowledge do LW readers have?
16. Book review: The Technology Trap
17. Wolf's Dice
18. What does Optimization Mean, Again? (Optimizing and Goodhart Effects - Clarifying Thoughts, Part 2)
19. June 2019 gwern.net newsletter
20. Why artificial optimism?
21. Doublecrux is for Building Products
22. What would be the signs of AI manhattan projects starting? Should a website be made watching for these signs?
23. Self-experiment Protocol: Effect of Chocolate on Sleep
24. The Right Way of Formulating a Problem?
25. Self-consciousness wants to make everything about itself
26. Largest open collection quotes about AI
27. Black hole narratives
28. Prereq: Cognitive Fusion
29. On the purposes of decision theory research
30. Job description for an independent AI alignment researcher
31. Diversify Your Friendship Portfolio
32. Should I wear wrist-weights while playing Beat Saber?
33. Dialogue on Appeals to Consequences
34. Why did we wait so long for the bicycle?
35. How often are new ideas discovered in old papers?
36. Let's Read: Superhuman AI for multiplayer poker
37. How to take smart notes (Ahrens, 2017)
38. Are we certain that gpt-2 and similar algorithms are not self-aware?
39. What are good resources for learning functional programming?
40. Intellectual Dark Matter
41. Where is the Meaning?
42. Is there neuroscience research on cognitive biases?
43. Prediction as coordination
44. Commentary On "The Abolition of Man"
45. Do bond yield curve inversions really indicate there is likely to be a recession?

46. An Increasingly Manipulative Newsfeed
47. If I knew how to make an omohundru optimizer, would I be able to do anything good with that knowledge?
48. Robust Agency for People and Organizations
49. Do you fear the rock or the hard place?
50. How to make a giant whiteboard for $14 (plus nails)

# Jeff Hawkins on neuromorphic AGI within 20 years

I just listened to AI podcast: Jeff Hawkins on the Thousand Brain Theory of Intelligence, and read some of the related papers. Jeff Hawkins is a theoretical neuroscientist; you may have heard of his 2004 book *On Intelligence*. Earlier, he had an illustrious career in EECS, including inventing the Palm Pilot. He now runs the company Numenta, which is dedicated to understanding how the human brain works (especially the neocortex), and using that knowledge to develop bio-inspired AI algorithms.

In no particular order, here are some highlights and commentary from the podcast and associated papers.

# Every part of the neocortex is running the same algorithm

The neocortex is the outermost and most evolutionarily-recent layer of the mammalian brain. In humans, it is about the size and shape of a dinner napkin (maybe 1500cm²×3mm), and constitutes 75% of the entire brain. Jeff wants us to think of it like 150,000 side-by-side "cortical columns", each of which is a little 1mm²×3mm tube, although I don't think we're supposed to the "column" thing too literally (there's no sharp demarcation between neighboring columns).

When you look at a diagram of the brain, the neocortex has loads of different parts that do different things—motor, sensory, visual, language, cognition, planning, and more. But Jeff says that all 150,000 of these cortical columns are virtually identical! Not only do they each have the same types of neurons, but they're laid out into the same configuration and wiring and larger-scale structures. In other words, there seems to be "general-purpose neocortical tissue", and if you dump visual information into it, it does visual processing, and if you connect it to motor control pathways, it does motor control, etc. He said that this theory originated with Vernon Mountcastle in the 1970s, and is now widely (but not universally) accepted in neuroscience. The theory is supported both by examining different parts of the brain under the microscope, and also by experiments, e.g. the fact that congenitally blind people can use their visual cortex for non-visual things, and conversely he mentioned in passing some old experiment where a scientist attached the optic nerve of a lemur to a different part of the cortex and it was able to see (or something like that).

Anyway, if you accept that premise, then there is one type of computation that the neocortex does, and if we can figure it out, we'll understand everything from how the brain does visual processing to how Einstein's brain invented General Relativity.

To me, cortical uniformity seems slightly at odds with the wide variety of instincts we have, like intuitive physics, intuitive biology, language, and so on. Are those not implemented in the neocortex? Are they implemented as connections between (rather than within) cortical columns? Or what? This didn't come up in the podcast. (ETA: I tried to answer this question in my later post, Human instincts, Symbol grounding, and the blank-slate neocortex.)

*(See also previous LW discussion at: The brain as a universal learning machine, 2015)*

# Grid cells and displacement cells

## Background: Grid cells for maps in the hippocampus

Grid cells, discovered in 2005, help animals build mental maps of physical spaces. (Grid cells are just one piece of a complicated machinery, along with "place cells" and other things, more on which shortly.) Grid cells are not traditionally associated with the neocortex, but rather the entorhinal cortex and hippocampus. But Jeff says that there's

some experimental evidence that they're also in the neocortex, and proposes that this is very important.

What are grid cells? Numenta has an educational video here. Here's my oversimplified 1D toy example (the modules can also be 2D). I have a cortical column with three "grid cell modules". One module consists of 9 neurons, one has 10 neurons, and the third has 11. As I stand in a certain position in a room, one neuron from each of the three modules is active - let's say the active neurons right now are $(x_1 \bmod 9)$, $(x_2 \bmod 10)$, and $(x_3 \bmod 11)$ for some integers $x_1, x_2, x_3$. When I take a step rightward, $x_1, x_2, x_3$ are each incremented by 1; when I take a step leftward, they're each decremented by 1. The three modules together can thus keep track of 990 unique spatial positions (cf. Chinese Remainder Theorem).

With enough grid cell modules of incommensurate size, scale-factor, and (in 2D) rotation, the number of unique representable positions becomes massive, and there is room to have lots of entirely different spaces (each with their own independent reference frame) stored this way without worrying about accidental collisions.

So you enter a new room. Your brain starts by picking a point in the room and assigns it a random $x_1, x_2, x_3$ (in my toy 1D example), and then stores all the other locations in the room in reference to that. Then you enter a hallway. As you turn your attention to this new space, you pick a *new* random $x_1, x_2, x_3$ and build your new hallway spatial map around there. So far so good, but there's a missing ingredient: the transformation from the room map to the hallway map, especially in their areas of overlap. How does that work? Jeff proposes (in this paper) that there exist what he calls "displacement cells", which (if I understand it correctly) literally implement modular arithmetic for the grid cell neurons in each grid cell module. So—still in the 1D toy example—the relation between the room map and the hall map might be represented by three displacement cell neurons $\delta_1, \delta_2, \delta_3$ (one for each of the three grid cell modules), and the neurons are wired up such that the brain can go back and forth between the activations

$$\{(x_1 \bmod 9), (x_2 \bmod 10), (x_3 \bmod 11)\} \quad \leftrightarrow$$

$$\leftrightarrow \quad \{((x_1 + \delta_1) \bmod 9), ((x_2 + \delta_2) \bmod 10), ((x_3 + \delta_3) \bmod 11)\}.$$

So if grid cell #2 is active, and then displacement cell #5 turns on, it should activate grid cell #7=5+2. It's kinda funny, but why not? We just put in a bunch of synapses that hardcode each entry of an addition table—and not even a particularly large one.

(Overall, all the stuff about the detailed mechanisms of grid cells and displacement cells comes across to me as "Ingenious workaround for the limitations of biology", not "Good idea that AI might want to copy", but maybe I'm missing something.)

# New idea: Grid cells for "maps" of objects and concepts in the neocortex

Anyway, Jeff theorizes that this grid cell machinery is not only used for navigating real spaces in the hippocampus but also navigating concept spaces in the neocortex.

*Example #1: A coffee cup.* We have a mental map of a coffee cup, and you can move around in that mental space by incrementing and decrementing the $x_i$ (in my 1D toy example).

*Example #2: A coffee mug with a picture on it.* Now, we have a mental map of the coffee mug, and a separate mental map of the picture, and then a set of displacement cells describe where the picture is in relation to the coffee cup. (This also includes relative rotation and scale, which I guess are also part of this grid cell + displacement cell machinery somehow, but he says he hasn't worked out all the details.)

*Example #3: A stapler*, where the two halves move with respect to each other. This motion can be described by a sequence of displacement cells ... and conveniently, neurons are excellent at learning temporal sequences (see below).

*Example #4: Logarithms*. Jeff thinks we have a reference frame for *everything*! Every word, every idea, every concept, everything you know has its own reference frame, in at least one of your cortical columns and probably thousands of them. Then displacement cells can encode the mathematical transformations of logarithms, and the relations between logarithms and other concepts, or something like that. I tried to sketch out an example of what he might be getting at in the next section below. Still, I found that his discussion of abstract cognition was a bit sketchier and more confusing than other things he talked about. My impression is that this is an aspect of the theory that he's still working on.

# "Thousand brains" theory

(See also Numenta educational video.) His idea here is that every one of the 150,000 "cortical columns" in the brain (see above) has the whole machinery with grid cells and displacement cells, reference frames for gazillions of different objects and concepts, and so on.

A cortical column that gets input from the tip of the finger is storing information and making predictions about what the tip of the finger will feel as it moves around the coffee cup. A cortical column in the visual cortex is storing information and making predictions about what it will *see* in *its* model of the coffee cup. And so on. If you reach into a box, and touch it with four fingers, each of those fingers is trying to fit its own data into its own model to learn what the object is, and there's a "voting" mechanism that allows them to reach agreement on what it is.

So I guess if you're doing a math problem with a logarithm, and you're visually imagining the word "log" floating to the other side of the equation and turning into an "exp", then there's a cortical column in your *visual* cortex that "knows" (temporal sequence memory) how this particular mathematical transformation works. Maybe the other cortical columns don't "know" that that transformation is possible, but can find out the result via the voting mechanism.

Or maybe you're doing the same math problem, but instead of visualizing the transformation, instead you recite to yourself the poem: "Inverse of log is exp". Well, then this knowledge is encoded as the temporal sequence memory in some cortical column of your *auditory* cortex.

There's a homunculus-esque intuition that all these hundreds of thousands of models need to be brought together into one unified world model. Neuroscientists calls this the "sensor fusion" problem. Jeff denies the whole premise. Thousands of different incomplete world models, plus a voting mechanism, is all you need; there is no unified world model.

Is the separate world model for each cortical column an "Ingenious workaround for the limitations of biology" or a "Good idea that AI should copy"? On the one hand, clearly there's *some* map between the concepts in different cortical columns, so that voting can work. That suggests that we can improve on biology by having one unified world model, but with many different coordinate systems and types of sensory prediction associated with each entry. On the other hand, maybe the map between entries of different columns' world models is *not* a nice one-to-one map, but rather some fuzzy many-to-many map. Then unifying it into a single ontology might be fundamentally impossible (except trivially, as a disjoint union). I'm not sure. I guess I should look up how the voting mechanism is supposed to work.

# Human-level AI, timelines, and existential risk

Jeff's goal is to "understand intelligence" and then use it to build intelligent machines. He is confident that this is possible, and that the machines can be dramatically smarter than humans (e.g. thinking faster, more memory, better at physics and math, etc.). Jeff thinks the hard part is done—he has the right framework for understanding cortical algorithms, even if there are still some details to be filled in. Thus, Jeff believes that, if he succeeds at proselytizing his understanding of brain algorithms to the AI community (which is why he was doing that podcast), then we should be able to make machines with human-like intelligence in less than 20 years.

Near the end of the podcast, Jeff emphatically denounced the idea of AI existential risk, or more generally that there was any reason to second-guess his mission of getting beyond-human-level intelligence as soon as possible. However, he appears to be profoundly misinformed about both what the arguments are for existential risk and who is making them. Ditto for Lex, the podcast host.

# Differences between actual neurons and artifical neural networks (ANNs)

## Non-proximal synapses and recognizing time-based patterns

He brought up his paper Why do neurons have thousands of synapses?. Neurons have anywhere from 5 to 30,000 synapses. There are two types. The synapses near the cell body (perhaps a few hundred) can cause the neuron to fire, and these are most similar to the connections in ANNs. The other 95% are way out on a dendrite (neuron branch), too far from the neuron body to make it fire, *even if all 95% were activated at once*! Instead, what happens is if you have 10-40 of these synapses that all activate at the same time *and are all very close to each other on the dendrite*, it creates a "dendritic

spike" that goes to the cell body and raises the voltage a little bit, but not enough to make the cell fire. And then the voltage goes back down shortly thereafter. What good is that? If the neuron is triggered to fire (due to the first type of synapses, the ones near the cell body), and has already been prepared by a dendritic spike, then it fires slightly *sooner*, which matters because there are fast inhibitory processes, such that if a neuron fires slightly before its neighbors, it can prevent those neighbors from firing at all.

So, there are dozens to hundreds of different patterns that the neuron can recognize—one for each close-together group of synapses on a dendrite—each of which can cause a dendritic spike. This allows networks of neurons to do sophisticated temporal predictions, he says: "Real neurons in the brain are time-based prediction engines, and there's no concept of this at all" in ANNs; "I don't think you can build intelligence without them".

Another nice thing about this is that a neuron can learn a new pattern by forming a new cluster of synapses out on some dendrite, and it won't alter the neuron's other behavior—i.e., it's an OR gate, so when that particular pattern is not occurring, the neuron behaves exactly as before.

# Binary weights, sparse representations

Another difference: "synapses are very unreliable"; you can't even assign one digit of precision to their connection strength. You have to think of it as almost binary. By contrast, I think most ANN weights are stored with at least ~2 and more often 7 decimal digits of precision.

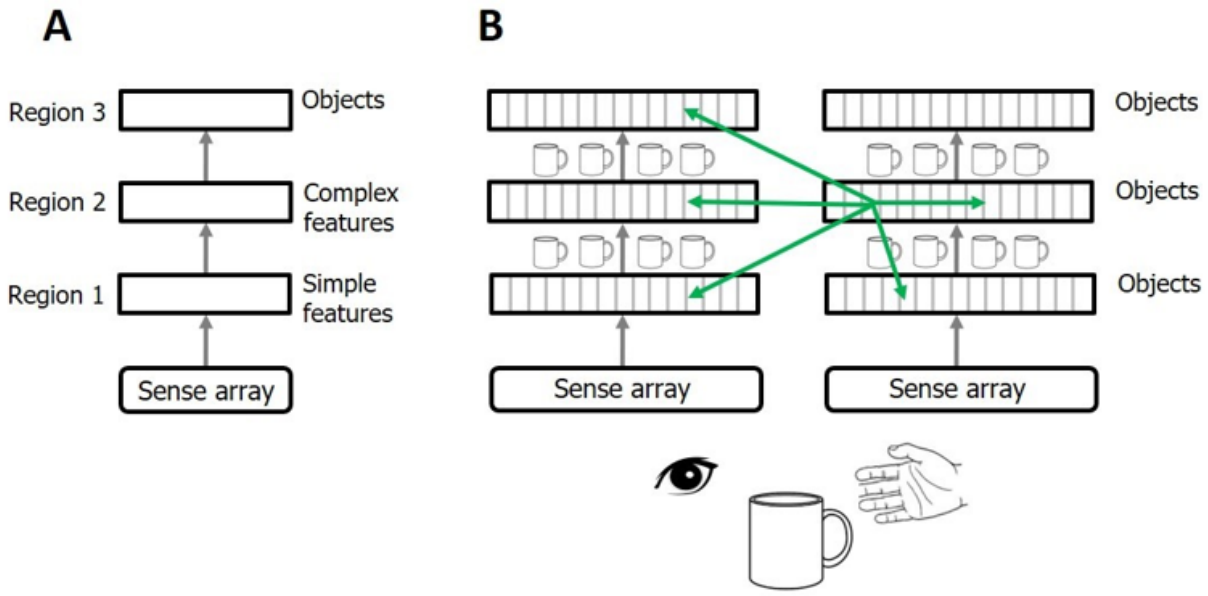
Related to this, "the brain works on sparse patterns". He mentioned his paper How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. He came back to this a couple times. Apparently in the brain, at any given moment, ~2% of neurons are firing. So imagine a little subpopulation of 10,000 neurons, and you're trying to represent something with a population code of sets of 200 of these neurons. First, there's an enormous space of possibilities ($10^{424}$). Second, if you pick two random sets-of-200, their overlap is almost always just a few. Even if you pick millions of sets, there won't be any pair that significantly overlaps. Therefore a neuron can "listen" for, say, 15 of the 200 neurons comprising X, and if those 15 all fire at once, that must have been X. The low overlap between different sets also gives the system robustness, for example to neuron death. Based on these ideas, they recently published this paper advocating for sparseness in image classifier networks, which sounds to me like they're reinventing neural network pruning, but maybe it's slightly different, or at least better motivated.

# Learning and synaptogenesis

According to Jeff, the brain does not learn by changing the strength of synapses, but rather by forming new synapses (synaptogenesis). Synaptogenesis takes about an hour. How does short-term memory work faster than that? There's something called "silent synapses", which are synapses that don't release neurotransmitters. Jeff's (unproven) theory is that short-term memory entails the conversion of silent synapses into active synapses, and that this occurs near-instantaneously.

# Vision processing

His [most recent paper] has this image of image processing in the visual cortex:




As I understand it, the idea is that every part of the field of view is trying to fit what it's looking at into its own world model. In other words, when you look at a cup, you shouldn't be thinking that the left, center, and right parts of the field-of-view are combined together and then the whole thing is recognized as a coffee cup, but rather that the left part of the field-of-view figures out that it's looking at the left side of the coffee cup, the center part of the field-of-view figures out that it's looking at the center of the coffee cup, and the right part of the field-of-view figures out that it's looking at the right side of the coffee cup. This process is facilitated by information exchange between different parts of the field-of-view, as well as integrating the information that a single cortical column sees over time as the eye or coffee cup moves. As evidence, they note that there are loads of connections in the visual cortex that are non-hierarchical (green arrows). Meanwhile, the different visual areas (V1, V2, etc.) are supposed to operate on different spatial scales, such that a faraway cup of coffee (taking up a tiny section of your field-of-view) might be recognized mainly in V1, while a close-up cup of coffee (taking up a larger chunk of your field-of-view) might be recognized mainly in V4, or something like that.

Maybe this has some profound implications for building CNN image classifiers, but I haven't figured out what exactly they would be, other than "Maybe try putting in a bunch of recurrent, non-hierarchical, and/or feedback connections?"

# My conclusions for AGI safety

Jeff's proud pursuit of superintelligence-as-fast-as-possible is a nice reminder that, despite the mainstreaming of AGI safety over the past few years, there's still a lot more advocacy and outreach work to be done. Again, I'm concerned not so much about the fact that he disagrees with arguments for AGI existential risks, but rather that he

(apparently) has never even *heard* the arguments for AGI existential risks, at least not from any source capable of explaining them correctly.

As for paths and timelines: I'm not in a great position to judge whether Jeff is on the right track, and there are way too many people who claim to understand the secrets of the brain for me to put a lot of weight on any one of them being profoundly correct. Still, I was pretty impressed, and I'm updating slightly in favor of neuromorphic AGI happening soon, particularly because of his claim that the whole neocortex is more-or-less cytoarchitecturally uniform.

Finally, maybe the most useful thing I got out of this is fleshing out my thinking about what an AGI's world model might look like.

Jeff is proposing that our human brain's world models are *ridiculously* profligate in the number of primitive entries included. Our world models don't just have one entry for "shirt", but rather separate entries for wet shirt, folded shirt, shirt-on-ironing-board, shirt-on-floor, shirt-on-our-body, shirt-on-someone-else's-body, etc. etc. etc. After all, each of those things is associated with a different suite of sensory predictions! In fact, it's even more profligate than that: Really, there might be an entry for "shirt on the floor, as it feels to the center part of my left insole when I step on it", and an entry for "my yellow T-shirt on the floor, as it appears to the rod cells in my right eye's upper peripheral vision". Likewise, instead of *one* entry for the word "shirt", there are thousands of them in the various columns of the auditory cortex (for the spoken word), and thousands more in the columns of the visual cortex (for the written word). To the extent that there's *any* generic abstract concept of "shirt" in the human brain, it would probably be some meta-level web of learned connections and associations and transformations between all these different entries.

If we build an AI which, like the human brain, has literally trillions of primitive elements in its world model, it seems hopeless to try to peer inside and interpret what it's thinking. But maybe it's not so bad? Let's say some part of cortical column #127360 has 2000 active neurons at some moment. We can break that down into 10 simultaneous active concepts (implemented as sparse population codes of 200 neurons each), and then for each of those 10, we can look back at the record of what was going on the first time that code ever appeared. We can look at the connections between that code and columns of the language center, and write down all those words. We can look at the connections between that code and columns of the visual cortex, and display all those images. Probably we can figure out more-or-less what that code is referring to, right? But it might take 1000 person-years to interpret one second of thought by a human-brain-like AGI! (...Unless we have access to an army of AI helpers, says the disembodied voice of Paul Christiano....) Also, some entries of the world model might be just plain illegible despite our best efforts, e.g. the various neural codes active in Ed Witten's brain when he thinks about theoretical physics.

# The Costs of Reliability

A question that used to puzzle me is “Why can people be so much better at doing a thing for fun, or to help their friends and family, than they are at doing the *exact same thing* as a job?”

I’ve seen it in myself and I’ve seen it in others. People can be hugely more productive, creative, intelligent, and efficient on just-for-fun stuff than they are at work.

Maybe it’s something around coercion? But it happens to people even when they choose their work and have no direct supervisor, as when a prolific hobbyist writer suddenly gets writer’s block as soon as he goes pro.

I think it has a very mundane explanation; it’s always more expensive to have to *meet a specific commitment* than merely to *do something valuable*.

If I feel like writing sometimes and not other times, then if writing is my hobby I’ll write when I feel like it, and my output per hour of writing will be fairly high. Even within “writing”, if my interests vary, and I write about whatever I feel like, I can take full advantage of every writing hour.  By contrast, if I’ve committed to write a specific piece by a specific deadline, I have to do it whether or not I’m in the mood for it, and that means I’ll probably be less efficient, spend more time dithering, and I’ll demand more external compensation in exchange for my inconvenience.

The stuff I write for fun may be valuable! And if you simply divide the value I produce by my hours of labor or the amount I need to be paid, I’m *hugely* more efficient in my free time than in my paid time! But I can’t just trivially “up my efficiency” in my paid time; reliability itself has a cost.

The costs of reliability are often invisible, but they can be very important. The cost (in time and in office supplies and software tools) of tracking and documenting your work so that you can deliver it on time. The cost (in labor and equipment) of quality assurance testing. The opportunity cost of creating simpler and less ambitious things so that you can deliver them on time and free of defects.

Reliability becomes more important with scale. Large organizations have more rules and procedures than small ones, and this is rational. Accordingly, they pay more costs in reliability.

One reason is that the attack surface for errors grows with the number of individuals involved. For instance, large organizations often have rules against downloading software onto company computers without permission.  The chance that any one person downloads malicious software that seriously harms the company is small, but the chance that *at least one* person does rises with the number of employees.

Another reason is that coordination becomes more important with more people. If a project depends on many people cooperating, then you as an individual aren’t simply trying to do the *best* thing, but rather the best thing that is also understandable and predictable and capable of achieving buy-in from others.

Finally, large institutions are more tempting to attackers than small ones, since they have more value to capture. For instance, large companies are more likely to be targeted by lawsuits or public outcry than private individuals, so it’s strategically

correct for them to spend more on defensive measures like legal compliance procedures or professional PR.

All of these types of defensive or preventative activity reduce efficiency — you can do less in a given timeframe with a given budget. Large institutions, even when doing everything right, acquire inefficiencies they didn't have when small, because they have higher reliability requirements.

Of course, there are also economies of scale that increase efficiency. There are fixed expenses that only large institutions can afford, that make marginal production cheaper. There are ways to aggregate many risky components so that the whole is *more* robust than any one part, e.g. in distributed computation, compressed sensing, or simply averaging. Optimal firm size is a balance.

This framework tells us when we ought to find it possible to get better-than-default efficiency *easily*, i.e. without any clever tricks, just by accepting different tradeoffs than others do. For example:

1.) People given an open-ended mandate to do what they like can be far more efficient than people working to spec…*at the cost of* unpredictable output with no guarantees of getting what you need when you need it. (See: academic research.)

2.) Things that come with fewer guarantees of reliable performance can be cheaper in the average use case…*at the cost of* completely letting you down when they occasionally fail. (See: prototype or beta-version technology.)

3.) Activities within highly cooperative social environments can be more efficient…*at the cost of* not scaling to more adversarial environments where you have to spend more resources on defending against attacks. (See: Eternal September)

4.) Having an "opportunistic" policy of taking whatever opportunities come along (for instance, hanging out in a public place and chatting with whomever comes along and seems interesting, vs. scheduling appointments) allows you to make use of time that others have to spend doing "nothing" … *at the cost of* never being able to commit to activities that need time blocked out in advance.

5.) Sharing high-fixed-cost items (like cars) can be more efficient than owning…*at the cost of* not having a guarantee that they'll always be available when you need them.

In general, you can get greater efficiency for things you don't *absolutely need* than for things you do; if something is merely nice-to-have, you can handle it if it occasionally fails, and your average cost-benefit ratio can be very good indeed. But this doesn't mean you can easily copy the efficiency of luxuries in the production of necessities.

(This suggests that "toys" are a good place to look for innovation. Frivolous, optional goods are where we should expect it to be most affordable to experiment, all else being equal; and we should expect technologies that first succeed in "toy" domains to expand to "important, life-and-death" domains later.)

# Integrity and accountability are core parts of rationality

*Epistemic Status: Pointing at early stage concepts, but with high confidence that something real is here. Hopefully not the final version of this post.*

When I started studying rationality and philosophy, I had the perspective that people who were in positions of power and influence should primarily focus on how to make good decisions in general and that we should generally give power to people who have demonstrated a good track record of general rationality. I also thought of power as this mostly unconstrained resource, similar to having money in your bank account, and that we should make sure to primarily allocate power to the people who are good at thinking and making decisions.

That picture has changed a lot over the years. While I think there is still a lot of value in the idea of "philosopher kings", I've made a variety of updates that significantly changed my relationship to allocating power in this way:

- I have come to believe that people's ability to come to correct opinions about important questions is in large part a result of whether their social and monetary incentives reward them when they have accurate models in a specific domain. This means a person can have extremely good opinions in one domain of reality, because they are subject to good incentives, while having highly inaccurate models in a large variety of other domains in which their incentives are not well optimized.
- People's rationality is much more defined by their ability to maneuver themselves into environments in which their external incentives align with their goals, than by their ability to have correct opinions while being subject to incentives they don't endorse. This is a tractable intervention and so the best people will be able to have vastly more accurate beliefs than the average person, but it means that "having accurate beliefs in one domain" doesn't straightforwardly generalize to "will have accurate beliefs in other domains".

  One *is* strongly predictive of the other, and that's in part due to general thinking skills and broad cognitive ability. But another major piece of the puzzle is the person's ability to build and seek out environments with good incentive structures.
- Everyone is highly irrational in their beliefs about at least some aspects of reality, and positions of power in particular tend to encourage strong incentives that don't tend to be optimally aligned with the truth. This means that highly competent people in positions of power often have less accurate beliefs than competent people who are not in positions of power.
- The design of systems that hold people who have power and influence accountable in a way that aligns their interests with both forming accurate beliefs and the interests of humanity at large is a really important problem, and is a major determinant of the overall quality of the decision-making ability of a community. General rationality training helps, but for collective decision making the creation of accountability systems, the tracking of outcome metrics and the design of incentives is at least as big of a factor as the degree to which the individual members of the community are able to come to accurate beliefs on their own.

A lot of these updates have also shaped my thinking while working at CEA, LessWrong and the LTF-Fund over the past 4 years. I've been in various positions of power, and have interacted with many people who had lots of power over the EA and Rationality communities, and I've become a lot more convinced that there is a lot of low-hanging fruit and important experimentation to be done to ensure better levels of accountability and incentive-design for the institutions that guide our community.

I also generally have broadly libertarian intuitions, and a lot of my ideas about how to build functional organizations are based on a more start-up like approach that is favored here in Silicon Valley. Initially these intuitions seemed at conflict with the intuitions for more emphasis on accountability structures, with broken legal systems, ad-hoc legislation, dysfunctional boards and dysfunctional institutions all coming to mind immediately as accountability-systems run wild. I've since then reconciled my thoughts on these topics a good bit.

# Integrity

Somewhat surprisingly, "integrity" has not been much discussed as a concept handle on LessWrong. But I've found it to be a pretty valuable virtue to meditate and reflect on.

I think of integrity as a more advanced form of honesty – when I say “integrity” I mean something like “acting in accordance with your stated beliefs.” Where honesty is the commitment to not speak direct falsehoods, integrity is the commitment to speak truths that actually ring true to yourself, not ones that are just abstractly defensible to other people. It is also a commitment to act on the truths that you do believe, and to communicate to others what your true beliefs are.

Integrity can be a double-edged sword. While it is good to judge people by the standards they expressed, it is also a surefire way to make people overly hesitant to update. If you get punished every time you change your mind because your new actions are now incongruent with the principles you explained to others before you changed your mind, then you are likely to stick with your principles for far longer than you would otherwise, even when evidence against your position is mounting.

The great benefit that I experienced from thinking of integrity as a virtue, is that it encourages me to build accurate models of my own mind and motivations. I can only act in line with ethical principles that are actually related to the real motivators of my actions. If I pretend to hold ethical principles that do not correspond to my motivators, then sooner or later my actions will diverge from my principles. I've come to think of a key part of integrity being the art of making accurate predictions about my own actions and communicating those as clearly as possible.

There are two natural ways to ensure that your stated principles are in line with your actions. You either adjust your stated principles until they match up with your actions, or you adjust your behavior to be in line with your stated principles. Both of those can backfire, and both of those can have significant positive effects.

# Who Should You Be Accountable To?

In the context of incentive design, I find thinking about integrity valuable because it feels to me like the natural complement to accountability. The purpose of

accountability is to ensure that you do what you say you are going to do, and integrity is the corresponding virtue of holding up well under high levels of accountability.

Highlighting accountability as a variable also highlights one of the biggest error modes of accountability and integrity – choosing too broad of an audience to hold yourself accountable to.

There is tradeoff between the size of the group that you are being held accountable by, and the complexity of the ethical principles you can act under. Too large of an audience, and you will be held accountable by the lowest common denominator of your values, which will rarely align well with what you actually think is moral (if you've done any kind of real reflection on moral principles).

Too *small* or too memetically close of an audience, and you risk not enough people paying attention to what you do, to actually help you notice inconsistencies in your stated beliefs and actions. And, the smaller the group that is holding you accountable is, the smaller your inner circle of trust, which reduces the amount of total resources that can be coordinated under your shared principles.

I think a major mistake that even many well-intentioned organizations make is to try to be held accountable by some vague conception of "the public". As they make public statements, someone in the public will misunderstand them, causing a spiral of less communication, resulting in more misunderstandings, resulting in even less communication, culminating into an organization that is completely opaque about any of its actions and intentions, with the only communication being filtered by a PR department that has little interest in the observers acquiring any beliefs that resemble reality.

I think a generally better setup is to choose a much smaller group of people that you trust to evaluate your actions very closely, and ideally do so in a way that is itself transparent to a broader audience. Common versions of this are auditors, as well as nonprofit boards that try to ensure the integrity of an organization.

This is all part of a broader reflection on trying to create good incentives for myself and the LessWrong team. I will try to follow this up with a post that more concretely summarizes my thoughts on how all of this applies to LessWrong concretely.

# In summary:

- One lens to view integrity through is as an advanced form of honesty – “acting in accordance with your stated beliefs.”
    - To improve integrity, you can either try to bring your actions in line with your stated beliefs, or your stated beliefs in line with your actions, or reworking both at the same time. These options all have failure modes, but potential benefits.
- People with power sometimes have incentives that systematically warp their ability to form accurate beliefs, and (correspondingly) to act with integrity.
- An important tool for maintaining integrity (in general, and in particular as you gain power) is to carefully think about what social environment and incentive structures you want for yourself.
- Choose carefully who, and how many people, you are accountable to:
    - Too many people, and you are limited in the complexity of the beliefs and actions that you can justify.

    - Too few people, too similar to you, and you won't have enough opportunities for people to notice and point out what you're doing wrong. You may also not end up with a strong enough coalition aligned with your principles to accomplish your goals.

[This post was originally posted on my shortform feed]

# A Key Power of the President is to Coordinate the Execution of Existing Concrete Plans

I listened to the 80,000 Hours podcast with Tom Kalil, who spent 16 years as Deputy Director of the Office of Science and Technology Policy at the White House. Kalil seems skilled at evaluating concrete scientific plans to offer the president and finding the path of least resistance through government to effect those plans, though he is not himself someone with deep technical understanding of any single domain.

One key idea I took from the podcast was that his main use of the executive branch of government is as a coordination mechanism. I moved away from thinking of the President as an expert who makes decisions like a CEO, and much more as an individual with immense coordination power trying his best to take any concrete plans given to him and coordinate the country around executing on them. That is, not someone who comes up with plans, not someone who executes on the plans, but someone who coordinates people to execute the concrete plans that are waiting to be picked up and run with.

Below are relevant and very interesting quotes, followed by a few more updates I made listening to the podcast.

## Key Quotes

> **Robert Wiblin:** So, do you think people under-appreciate how much the executive branch can just do autonomously?
>
> **Tom Kalil:** Yes. Yeah. Not only what it can do, but the President's ability to convene.
>
> **[...]**
>
> **Tom Kalil:** One thing that I used to ask people is to imagine that you have a 15 minute meeting with the President in the Oval Office and he says, "Rob, if you give me a good idea for", pick your cause, "reducing existential risk, then I will call anyone on the planet. It can be a conference call so there can be more than one person on the line. If it's someone from inside the government, that I can direct them to something because I'm their boss, and if it's someone outside the government then I can challenge them to do something. So, you not only have to tell me, what is your idea, but in order to make your idea happen, who would I call and what would I ask them to do?"
>
> **Tom Kalil:** There are several reasons for this thought experiment. One is that if you work for the President you have the ability to send the President a decision memo and have him check the box that says yes. Over time that give you a sense of what psychologists call agency, a sense that many things that you see in the world around you are the result of human action or inaction, as opposed to the laws of physics. That's one thing, a more expansive view of what do you think is potentially changeable. The second is, it's sort of a version of the Hamming

question, presumably if you really did have a meeting with the President you'd use it to describe an issue that you thought was really important as opposed to a secondary or third tier issue. The third is that many complex problems cannot be solved by a single individual organization, they require coalitions.

**Tom Kalil:** You can't build a coalition if you can't articulate, number one, who are the members of the coalition, and number two, what are the mutually reinforcing steps that you would want them to take. That's one thing that I talk about in the Policy Entrepreneurship, then I also talk about something that people don't ever really appreciate, which is that policy makers do things with words. What do I mean by that? Well, think about when the priest says, "I now pronounce you man and wife", he has changed the state of affairs by virtue of, A, him being a priest, and B, him saying, "I now pronounce you man and wife". Similarly, the way that a policy maker both frames and makes a decision and implements that decision is through documents. When the President does it we call it an executive order or presidential memorandum, when a regulatory agency does it we call it a rule, when the Congress does it we call it legislation. But in all instances it is a document that you are creating or editing, so part of the policy process is that you are able to figure out what's the document or documents that you need to create or edit and who is allowed to take that something from being a Word document that is on your screen to something that has some force in the world?

**Tom Kalil:** I would see this all the time, something would go from being a Word document on my computer to being a presidential executive order, it always seemed like this slightly magical transformation from a Word Doc to something that is instructing relevant members of the Cabinet to take some action.

**Robert Wiblin:** I guess this makes you more ambitious, then you're like, "What is the best thing, what is the best memo that I could write.

**Tom Kalil:** Yeah, exactly, yeah. But also you have to be able to articulate some coherent relationship between ends and means. I would … a lot of times someone would come visit me and they would say, "my issue is important".

**Tom Kalil:** I'd say, "great, let's say that I'm prepared to stipulate that, what is that you want me to do?", then they would look at me and they would say, "Well, you should make this a priority".

**Tom Kalil:** I'd say, "What would that look like?" People were not able … they were able to tell you that their issue was important and that they thought the President should devote more time and energy to it, but then when you said "Alright, what is it, literally … let's say we got the President super interested in this issue, what would they do?" They weren't able to articulate that part.

**Robert Wiblin:** Yeah, I'm sympathetic to that 'cause there's a lot of things that I think are very important, but I'm also not sure what should be done. I suppose maybe it makes sense for people to think more about that once they've gotten people to care about it, but at the same time, maybe it's hard to get people to care about something if you have no actual concrete steps that they can take, they're like, "Well, I don't know what to do".

**Tom Kalil:** Yes. Yeah, because that's assuming that … because what you're saying is, I've really thought about this issue a lot and I think it's really important, but I don't know what to do."

**Robert Wiblin:** That's a bad sign. So you should think about it.

Kalil also talks about in his role as Deputy Director of the Office of Science and Technology Policy, he helped raise the staff count from 40 to over 100 during the Obama administration. He just gets to hire people who are excited about an idea and want to make it happen, and then they make it happen using the coordination power of the executive office. Here's a prominent example:

**Tom Kalil:** Let me give you one example. A young woman emailed me and the subject line of your email was, "Cass Sunstein says I should work for you."

**Robert Wiblin:** That's a strong subject line.

**Tom Kalil:** Good subject line. So I did a little research on her. It turned out that she had been a child violin prodigy with Itzhak Perlman, had won the major Yale undergraduate awards, was a Rhodes scholar, and was wrapping up a post-doc at Stanford in Decision Neuroscience. I went out on a limb and I decided to take a chance on her. Her name was Maya Shankar. I asked Maya, "What do you want to do?"

**Tom Kalil:** She said, "The UK has created this organization called the Behavioral Insights Team, which is taking these insights from people like Kahneman and Tversky and Sunstein and Thaler and using them to inform policies and programs. These are all US researchers. Why don't we have something like this?" She said, "I would like to create that."

**Tom Kalil:** Sure enough, in her late twenties, she arrived with no money, created this new organization called the Social and Behavioral Sciences Team, recruited 20 behavioral scientists to the federal government, got them to launch 60 collaborations with federal departments and agencies and got President Obama to sign an executive order institutionalizing this new entity.

**Tom Kalil:** I think that's pretty consequential for someone in their late twenties to be able to accomplish. That's one thing I did, was to recruit people of that caliber and teach them how to get things done in the federal government because the government doesn't come with an operating manual.

The next quote is about how the core goal of the office of science and technology policy is to take the necessary steps to get the private sector to build new tech:

**Tom Kalil:** One of the things I learned is that if the United States is behind in a technology, it's very difficult to try to re-establish a leadership position… We tried to do that in the area of technologies like flat panel displays and we invested some money, but I don't think a whole lot came out of it.

**Tom Kalil:** Once Korea and Japan dominated the market for things like active matrix, liquid crystal displays, then trying to get the United States back into that market is really, really hard, and might require more money than the US is willing to put into it. Because obviously we believe that the primary role of government is to create the right environment for the private sector. It's not to engage in this sort of heavy handed-top down industrial policy that you see a China engaging in, for example… we invested in this idea of flexible electronics where the idea is – maybe you have a display that's a piece of paper that you can roll up and put into your pocket. And, if that's an area where no one has established a clear leadership

position, that's more likely to be effective than saying, okay, we're going to duke it out in some market that we've kind of already lost.

This final quote is an example of the coordination power of the President.

**Robert Wiblin:** Having worked in White House for 16 years, you must have some interesting or funny stories from your experiences there, different perhaps than what people expect? Can you share one of them?

**Tom Kalil:** Sure. This is a story that happened in 1995 and 1996, so as I mentioned, Vice President Gore was really interested in this idea of the information superhighway and one of his goals was, what if we could connect every classroom to the internet? So I would tell people about the vice president's interest in this issue and someone who I'd gotten to know, John Gage, who was at a company by the name of Sun Microsystems said, "Oh, I've got this idea called Net Day. The idea is, what if on a single day, tens of thousands of engineers showed up in schools all across California and started the process of wiring California classrooms to the internet?" I said, "Well great." He said, "You know, I've got a web page of what this would look like if it actually happened." So he emailed it to me and I gave it to the vice president and the vice president thought it was a done deal.

**Tom Kalil:** So at this point, it was just in the fevered imagination of John Gage. So the vice president has a weekly lunch with the president and so he said, "Mr. President, we have Sun, we have Apple, we have HP, we have IBM, we have Pacific Gas and Electric and they have all agreed that they're going to wire thousands of classrooms and schools all across California." The president was like, "Great. Let's announce it." So it turned out that they were going to be in the Bay Area anyway so they decided, we're going to announce Net Day. So I called up John Gage and I said, "They're going to come out and announce this." So he and I spent the next week calling in every favor that we had to get these CEOs to show up and announce that they were for us. They were a little sketchy on the details of what it was.

**Tom Kalil:** What John did was he developed a website, which was a clickable map of California, that allowed you to zoom all the way down to the street level, all 12,000 public and private K through 12 schools had their own homepage. You could indicate your level of expertise from, "I am an experienced network engineer." To, "I will bring coffee and donuts." All the schools were color-coded red, yellow and green depending on how many volunteers had signed up. So we could look at the map and figure out which communities were getting onboard and which needed some positive reinforcement.

**Tom Kalil:** So they announced that not only were they supporting it but they were going to come back and personally participate in it. So by the time they did, we actually had tens of thousands of people who had volunteered, so it was this positive, self-fulfilling prophecy because they said, "Oh, there's going to be a Net Day." In fact, there was a Net Day and tens of thousands of engineers showed up to wire the schools and many parents showed up to wire the schools, but they discovered the windows were broken and the bathrooms didn't work, so a lot of them got more engaged in the schools as a result.

**Tom Kalil:** Many states decided they were going to do this and entire countries decided that they were going to have a Net Day as well. So it was this experience,

a couple of things that I took away from it, one is that you could create this positive self-fulfilling prophecy, even though that was a very nerve-wracking period of time for me personally because I'd committed to the president and vice president to do something-

**Robert Wiblin:** To announce the thing.

**Tom Kalil:** … and announce something that didn't really exist yet. Right?

**Robert Wiblin:** Yeah, sounds a little bit like an episode of Veep.

**Tom Kalil:** Then it was sort of applying massive parallels to this problem. So as opposed to saying, "How are we going to wire 10,000 schools?" The question was, how could you get every community to take responsibility for one school? So it was just very interesting of the experience that I had of going from something being a complete fantasy to actually seeing it happen.

**Robert Wiblin:** So government can get things done.

**Tom Kalil:** Yes.

**Robert Wiblin:** Just in sometimes a peculiar manner.

**Tom Kalil:** Yes, exactly.

There is a lot more fascinating discussion in the interview, especially Kalil's comments on using financial prizes to incentivise science+tech in areas like education and poverty.

# Updates

My new model is that the President's interaction with science is largely to take concrete ideas floating around in the environment that are ready for their time, and push them over the edge into actually being built by the US private sector, or into actually substantially informing government policy. This is similar to the notion that scientific ideas come about when the environment is ready for them (Newton and Leibniz both discovering calculus at the same time). There are executable plans floating around in the ether, and the President keeps getting handed them and sets them off. His department is not an originator of new ideas, it coordinates the execution of existing ones. (And there's a natural frame from which is this is the correct marginal use of attention from the President: compare 15 minutes per project versus spending a week becoming an expert in one and then executing it himself.)

I've updated positively on the tractability of gaining influence within the government and being able to use it on timescales of 4-8 years. (I expect I will likely make a further update when I read the blogposts of Dominic Cummings regarding UK politics, though not sure how strongly.) Overall I think influence in government, if you're ambitious and well-connected and have a very concrete vision, is likely quite a real action one can take. I expect that from the perspective of government there is a lot of low hanging fruit to be picked.

I updated negatively on the usefulness of interacting with this part of government in the short-to-medium term. My sense is that the state of understanding of how transformative AI will be built and what impact it will have on the world is sufficiently

low resolution and confused that we have no project or policy recommendations for the government, and will not be able to offer anything until we see further work that helps conceptualise this space. Listening to the podcast tells me that if you get 15 minutes to talk to the President about x-risk today, you are wasting his time, because we have no concrete plan that needs executing if only could coordinate major AI tech companies. We have no R&D projects that need funding. We have no nuanced AI-development policies for global powers to agree to. I'm pretty sure that there are people in this community who can coordinate Elon Musk and Demis Hassabis or whomever else, should we have an actionable plan, but the current state is that we have no plan to offer.

# Forum participation as a research strategy

Previously: Online discussion is better than pre-publication peer review, Disincentives for participating on LW/AF

Recently I've noticed a cognitive dissonance in myself, where I can see that my best ideas have come from participating on various mailing lists and forums (such as cypherpunks, extropians, SL4, everything-list, LessWrong and AI Alignment Forum), and I've received a certain amount of recognition as a result, but when someone asks me what I actually do as an "independent researcher", I'm embarrassed to say that I mostly comment on other people's posts, participate in online discussions, and occasionally a new idea pops into my head and I write it down as a blog/forum post of my own. I guess that's because I imagine it doesn't fit most people's image of what a researcher's work consists of.

Once I noticed this, the tension is easy to resolve - in this post I'm going to proclaim/endorse forum participation (aka commenting) as a productive research strategy that I've managed to stumble upon, and recommend it to others (at least to try). Note that this is different from saying that forum/blog posts are a good way for a research community to communicate. It's about individually doing better as researchers.

# Benefits of Forum Participation (FP)

## FP takes little effort / will power

In other words it feels more like play than work, which means I rarely have issues with not wanting to do something that I think is important to do (i.e., akrasia), the only exception being that writing posts seems to take more effort so occasionally I spend my time writing comments when I perhaps should write posts instead. (This is the part of this post that I think may be least likely to generalize to other people. It could be that I'm an extreme outlier in finding FP so low-effort. However it might also be the case that it becomes low effort for most people to write comments once they've had enough practice in it.)

## FP is a good way to notice missing background knowledge and provides incentives to learn missing knowledge

If you read a post with an intention to question or comment on it, it's pretty easy to notice that it assumes some background knowledge that you lack. The desire to not ask a "stupid" question or make a "stupid" comment provides powerful incentive to learn the miss knowledge.

## FP is a good way to stay up to date on everyone else's latest research

It's often a good idea to stay up to date on other people's research, but sometimes one isn't highly motivated to do so. FP seems to make that easier. For example, I wasn't following Stuart's research on counterfactual oracles, until the recent contest drew my attention and desire to participate, and I ended up reading the latest posts on CO in order to understand the current state of the art on that topic, which turned out to be pretty interesting.

## Arguments that are generated in reaction to some specific post or discussion can be of general value

It's not infrequent that I come up with an argument in response to some post or discussion thread, and later expand or follow up that argument into a post because it seems to apply more generally than to just that post/discussion. Here is one such example.

## FP generates new ideas via cross-fertilization

FP incentivizes one to think deeply about many threads of research, and often (at least for me) an idea pops into my head that seems to combine various partial ideas floating in the ether into a coherent or semi-coherent whole (e.g., UDT), or is the result of applying or analogizing someone else's latest idea to a different topic (e.g., "human safety problem", "philosophy as high complexity class").

## FP helps prepare for efficiently communicating new ideas

FP is a good way to build models of other people's epistemic states, and also a good way to practice communicating with fellow researchers, both of which are good preparation for efficiently communicating one's own new ideas.

# My Recommendations

## Comment more

To obtain the above benefits, one just has to write more comments. It may be necessary to first overcome disincentives to participate. If you can't, please speak up and maybe the forum admins will do something to help address whatever obstacle you're having trouble with.

## Practice makes better

If it seems hard to write good comments, practice might make it easier eventually.

## Think of FP as something to do for yourself

Some people might think of commenting as primarily providing a service to other researchers or to the research community. I suggest also thinking of it as providing a benefit to yourself (for the above reasons).

## Encourage and support researchers who adopt FP as their primary research strategy

I'm not aware of any organizations that explicitly encourage and support researchers to spend most or much of their time commenting on forum posts. But perhaps they should, if it actually is (or has the potential to be) a productive research strategy? For example this could be done by providing financial support and/or status rewards for effective forum participation.

# The Real Rules Have No Exceptions

*(This is a comment that has been turned into a post.)*

From Chris_Leong's post, "Making Exceptions to General Rules":

> Suppose you make a general rule, ie. "I won't eat any cookies". Then you encounter a situation that legitimately feels exceptional , "These are generally considered the best cookies in the entire state". This tends to make people torn between two threads of reasoning:
>
> 1. Clearly the optimal strategy is to make an exception this one time and then follow the rule the rest of the time.
>
> 2. If you break the rule this one time, then you risk dismantling the rule and ending up not following it at all.
>
> How can we resolve this? ...

This is my answer:

*Consider even a single exception to totally undermine any rule. Consequently, only follow rules with no exceptions.[1]. When you do encounter a legitimate exception to a heretofore-exceptionless rule, immediately discard the rule and replace it with a new rule—one which accounts for situations like this one, which, to the old rule, had to be exceptions.*

This, of course, requires a meta-rule (or, if you like, a meta-habit):

*Prefer simplicity in your rules. Be vigilant that your rules do not grow too complex; make sure you are not relaxing the legitimacy criteria of your exceptions. Periodically audit your rules, inspecting them for complexity; try to formulate simpler versions of complex rules.*

So, when you encounter an exception, you *neither* break the rule once but keep following it thereafter, *nor* break it once and risk breaking it again. If this is really an exception, then that rule is immediately and automatically nullified, because good rules ought not have exceptions. Time for a new rule.

And if you're not prepared to discard the rule and formulate a new one, well, then the exception must not be all that compelling; in which case, of course, keep following the existing rule, now and henceforth.

But why do I say that good rules ought not have exceptions? Because **rules already don't have exceptions**.

Exceptions are a fiction. They're a way for us to avoid admitting (sometimes to ourselves, sometimes to others) that the rule as stated, *together with the criteria for deciding whether something is a "legitimate" exception*, **is the actual rule**.

The approach I describe above merely consists of making this fact explicit.

---

1. By which I mean “only follow rules to which no legitimate exception will ever be encountered”, *not* “continue following a rule even if you encounter what seems like a legitimate exception”. ↩

# No nonsense version of the "racial algorithm bias"

In discussions of algorithm bias, the COMPAS scandal has been too often quoted out of context. This post gives the facts, and the interpretation, as quickly as possible. See this for details.

**The fight**

The COMPAS system is a statistical decision algorithm trained on past statistical data on American convicts. It takes as inputs features about the convict and outputs a "risk score" that indicates how likely the convict would reoffend if released.

In 2016, ProPublica organization claimed that COMPAS is clearly unfair for blacks in one way. Northpointe replied that it is approximately fair in another way. ProPublica rebukes with many statistical details that I didn't read.

The basic paradox at the heart of the contention is very simple and is not a simple "machines are biased because it learns from history and history is biased". It's just that there are many kinds of fairness, each may sound reasonable, but they are not compatible in realistic circumstances. Northpointe chose one and ProPublica chose another.

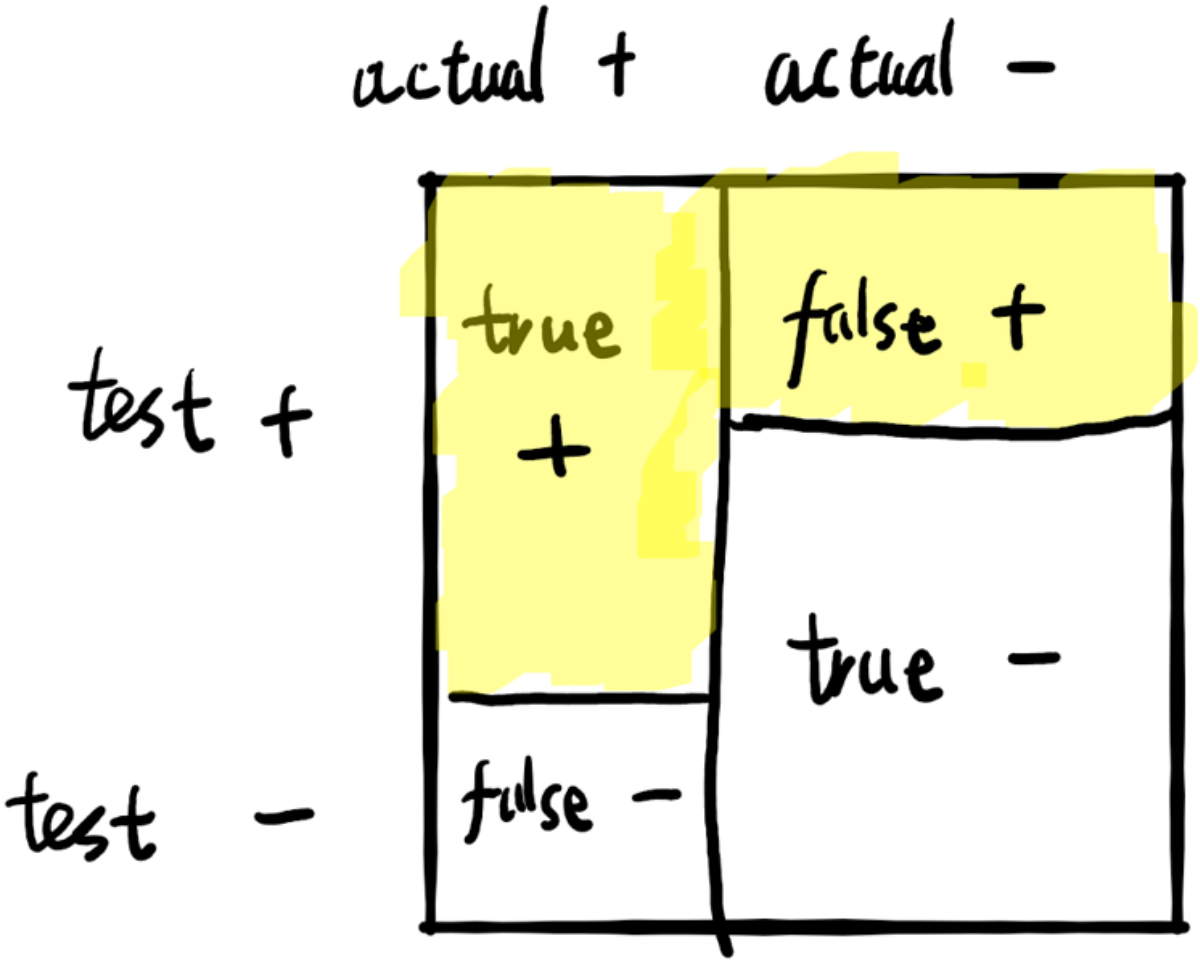
**The math**

The actual COMPAS gives a risk score from 1-10, but there's no need. Consider the toy example where we have a decider (COMPAS, a jury, or a judge) judging whether a group of convicts would reoffend or not. How well the decider is doing can be measured in at least three ways:

- False negative rate = (false negative)/(actual positive)
- False positive rate = (false positive)/(actual negative)
- Calibration = (true positive)/(test positive)

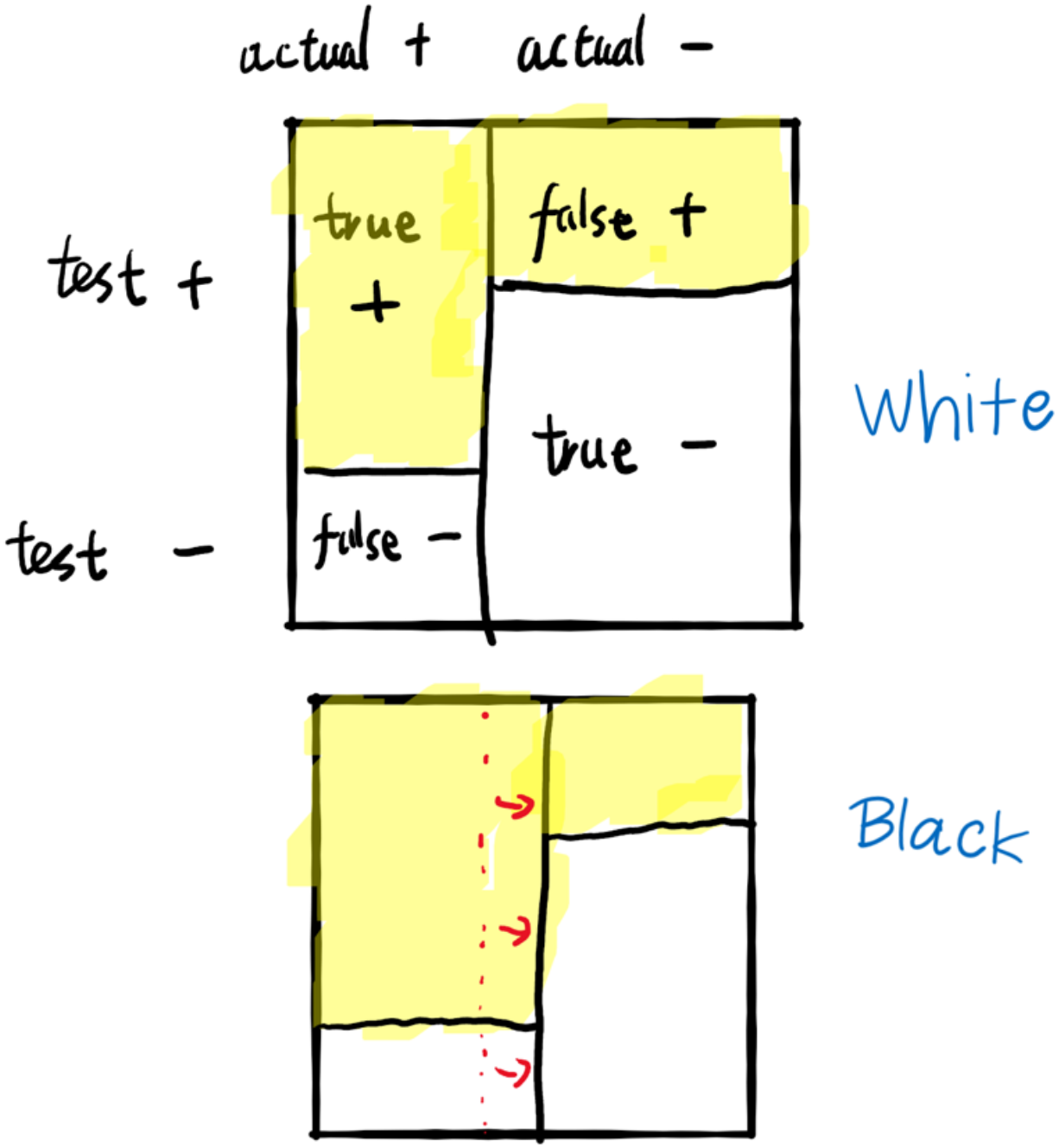
A good decider should have false negative rate close to 0, false positive rate close to 0, and calibration close to 1.

Visually, we can draw a "square" with four blocks:




- false negative rate = the "height" of the false negative block,
- false positive rate = the "height" of the false positive block,
- calibration = (true positive block)/(total area of the yellow blocks)

Now consider black convicts and white convicts. Now we have two squares. Since they have different reoffend rates for some reason, the central vertical line of the two squares are different.




The decider tries to be fair by making sure that the false negative rate and false positive rates are the same in both squares, but then it will be forced to make the calibration in the Whites lower than the calibration in the Blacks.

Then suppose the decider try to increase the calibration in the Whites, then the decider must somehow decrease the false negative rate of Whites, or the false positive rate of Whites.

In other words, when the base rates are different, it's impossible to have equal fairness measures in:

- false negative rate

- false positive rate
- calibration

Oh, forgot to mention, even when base rates are different, there's a way to have equal fairness measures in all three of those... But that requires the decider to be *perfect*: Its false positive rate and false negative rate must both be 0, and its calibration must be 1. This is unrealistic.

In the jargon of fairness measurement, "equal false negative rate and false positive rate" is "parity fairness"; "equal calibration" is just "calibration fairness". Parity fairness and calibration fairness can be straightforwardly generalized for COMPAS, which uses a 1-10 scoring scale, or indeed any numerical risk score.

It's some straightforward algebra to prove that in this general case, parity fairness and calibration fairness are incompatible when the base rates are different, and the decider is not perfect.

**The fight, after-math**

Northpointe showed that COMPAS is approximately fair in calibration for Whites and Blacks. ProPublica showed that COMPAS is unfair in parity.

The lesson is that there are incompatible fairnesses. To figure out which to apply -- that is a different question.

# Appeal to Consequence, Value Tensions, And Robust Organizations

**Epistemic Status: Strong opinions weakly held. Mostly trying to bring some things into the discourse that I think are too often ignored.**

*Some updates I've made based on the discussion in this post are here* .

# Introduction

Jessicata's Dialogue on Appeals to Consequences is an expansion of a response that she wrote to me a few months ago, arguing a particular point that I agree with: Namely, if you have an object level thing you want in the world, it's almost never worth lying or withholding information about that thing, because it breaks meta level norms about truthseeking that are much more important to accomplishing object level goals in general. However, there's a slightly more interesting case that I think is quite murkier, that the original comment was pointing to. That is, what if your truthseeking norms are in tension with OTHER meta level norms that are important? In general, how do you deal with instances where tensions between two important values cause you to not know what to do?

# Dialogue

Let's imagine John and Jill are discussing John's behavior in a private space. Jill is a leader of the space, and John is someone who frequently attends the space and has lively discussions trying to get to the truth.

Jill: John, I've had several complaints about your tendency to steer conversations towards the divisive topic that everyone should be a Vegan, and I'm going to ask you to tone it down a bit when you're in our main space.

John: Are people saying that I'm making arguments that are false?

Jill: No, no one is saying that you're making false arguments. John: Are people saying that I'm derailing the conversation? I think you'll find that every instance I brought up veganism was highly relevant to the conversation.

Jill: Yes, some people have said that, but I happen to believe you when you say that you've only brought it up in relevant contexts for you.

John: Then what's the problem? I'm stating relevant true beliefs that add to the totality of the conversation and steer it in conversationally relevant directions.

Jill: The problem is twofold. Firstly, people find it annoying to retread the same conversation over and over. More importantly, this topic usually leads to demon conversations, and I fear that continued discussion of the topic at the rate its' currently discussed could lead to a schism. Both of these outcomes go against our value of being a premiere community that attracts the smartest people, as they're actually driving these people away!

John: Excuse me for saying so, but this a clear appeal to consequence!

Jill: Is it? I'm not saying that the negative consequences to the community mean that what you're saying is false - that would be a clear logical fallacy. Instead I'm just asking you to bring up this argument less often because I think it will lead to bad outcomes.

John: Ok, maybe it's not a logical fallacy, but it is dangerous. This community is built on a foundation of truth seeking, and once we start abandoning that because of people's feelings, we devolve into tribal dynamics and tone arguments!

Jill: Yes, truthseeking is very important. However, It's clear that just choosing one value as sacred , and not allowing for tradeoffs can lead to very dysfunctional belief systems,.I believe you've pointed at a clear tension in our values as they're currently stated. The tension between freedom of speech and truth, and the value of making a space that people actually want to have intellectual discussions at.

John: You're saying there's a tension, but to me there's a clear and obvious winner. Under your proposed rules, anyone will be able to silence anything simply by saying they don't like it!

Jill: If I find someone trying to silence good arguments through that tactic, I'll sit them down and have a similar conversation to the one we're having now.

John: That's even worse! That means that instead of the putting the allowed conversation topics up to vote, we're putting them in the hands of one person, you! You can silence any conversation you want.

Jill: I can see how it would seem that way, but I believe we've cultivated some great cultural norms that make it harder for me to play to political games like that. Firstly, our norm of radical transparency means that this and all similar conversations I have like this will be recorded and shared with everyone, and any such political moves by me will be laughably transparent.

John: That makes sense. Also, Hi Mom!

Jill: Second, our organization allows anyone to apply the values to anyone else, so if you see ME not following the values in any of my talks, you can call me out on it and I'll comply.

John: Sure, you say that now, but because of your role you can just defy that rule whenever you want! Jill: That's true, and it's one of the reasons I've worked to cultivate integrity as a leader. Has there been any instance of my behavior where you think I would actually do that?

John: No I suppose not. Are there any other cultural norms preventing you from using the arbitrary nature of decisions for your own gain? Jill: There's one more. Our organization has a clear set of values, and as the leader one of my roles is to spearhead the change the values in clear ways when there's tension between them. So I'm not just going to talk to you, I'm actually going to suggest to the organization that we clarify our values such that they tell us to do in these relatively common situations, and I'm going to have you help me.

John: I think that makes sense. We can probably make a list of topics that people are allowed to taboo, and a list of topics people are not allowed to taboo, and then I'll

always know what it's ok to "appeal to consequences" on. Jill: I'm afraid that particular rule would be unwise. I think there's practically unlimited scissor statements that could cause schisms in our community, and a skilled adversary could easily find one that's not on our list of approved topics. No, I'm afraid we'll need to make a general value that can cover these situations in the general case.

John: Oh, so trying to avoid appeals to consequence argument can actually be used by someone looking to harm our community? That's interesting! But it's not clear to me that there is a general rule that can cover all the cases.

Jill: There is. The general rule is that people should give equal weight to their own needs, the needs of the people they're interacting with, and the needs of the organization as a whole.

John: I'm not sure I get it.

Jill: Well, you have a need to express that everyone should be a vegan. It's clearly very important to you, or you wouldn't bring it up so much. At the same time, many of the people in our community have a need to have variety in their conversation, and you should be aware of this when talking with them. Finally, our organization has a need to not experience/discuss scissor statements too often or too frequently, in order to remain healthy and avoid frequent schisms. By bringing this topic up so much, you're putting your needs above the needs of others you're interacting with and the group, instead of bringing it up less frequently, which would be placing the needs on equal ground.

John: That makes sense. I suppose by the same token, if there's a really interesting topic that's helpful for the group to know about, and lots of people want to talk about, it would be putting your own needs above others needs if you said it hurt your feelings so people couldn't talk about it.

Jill: Exactly!

John: So this rule seems plausible to me, and I'm sure it would be great for many people, but I have to admit its' not for me. I'd much prefer a space where people are allowed to say anything they want to me, and I can say anything I want to them in return.

Jill: I agree that this may not be the best rule for everybody. That's why next week we're going to start experimenting with The Archipelago Model. As I said, I want you to tone it down in the main room, which follows the Maturity value mentioned above. However, we've designated a side room that instead follows Crocker's Rules. You're allowed to go to either room, but when in that room, must follow the stated values of the room. And most importantly, all conversations are recorded and can be listened to by anyone in the community!

John: Cool, that seems worthwhile, but very messy and likely to have numerous hidden failure modes...

Jill: I agree, but it at least seems worth a shot!

# Commentary

So you probably noticed already, but this post wasn't really about Appeal to Consequences at all. Instead, it's a meditation on how good organizations deal with tensions in their values, and avoid the organization being overrun by skilled sociopaths. A lot of these suggestions and ideas come from the work I've been doing over the past year or so to figure out what makes great organizations and communities. I'd be particularly interested in peoples' inner sim of how the organization described by John and Jill above would go horribly wrong, and counter ideas about what could be done to fix THOSE issues.

# Everybody Knows

*“Everybody knows that the dice are loaded.*

*Everybody rolls with their fingers crossed.*

*Everybody knows the war is over.*

*Everybody knows the good guys lost.”*

– Leonard Cohen, Everybody Knows

“It is known.” – Dothraki saying

It is not known. Everybody doesn’t know.

When someone claims that everyone knows something, either they are short-cutting and specifically mean ‘everyone in this well-defined small group where complex common knowledge of this particular thing is something we have invested in,’ they are very wrong about how the world works, or much more commonly, they are *flat out lying.*

Saying that everybody knows is almost never a *mistake.* The statement isn’t sloppy reasoning. It’s a strategy that aims to cut off discussion or objection, to justify fraud and deception, and to establish truth without evidence.

## Not Everybody Knows

Let us first establish quickly that everyone doesn’t know. There are many ways to see this.

One way to see this is to point out that when Alice tells Bob that everybody knows X, either Bob is asserting X because people act as if they don’t know X, or *Bob does not know X.* That’s why Alice is telling Bob in the first place.

A second way is to attempt to explain something in detail as you would to a child.

A cleaner way is to consider some examples of things that a lot of people don’t know. According to the first Google hit, 32 million American adults can’t read, and 50% can’t read a book at the 8th grade level. Various other tests of basic skills from school don’t look much better. Here are some more basic facts many Americans don’t know, including 20% who think the Sun revolves around the Earth. Nigerian prince scams still make over $700,000 per year. Doctors can’t do basic job-relevant probability calculations within an order of magnitude. Just yesterday (as of writing this) I had to explain to a college graduate that Bitcoin was more volatile than the stock market, and Forex was not a responsible retirement savings plan.

## What does the claim that ‘everybody knows’ mean?

There are a few different things ‘everybody knows’ is standing in for when someone claims it.

In most of them, the claim that literal actual ‘everybody knows’ is sort of the Bailey, and the thing we’ll describe here is the implicit Motte that ‘everyone knows’ is your real message. Which of course, in turn, not everybody knows. As is often the case, *the Bailey is blatantly false.* But demonstrating that is socially costly. It shows you are the one who does not get it, who is not in on the goings on. So much so that when someone ‘calls someone out’ on a blatant lie, the liar socially benefits.

I see four related central modes. They overlap and reinforce each other, and are often all in play at once.

The first central mode is ‘this is obviously true because social proof, so I don’t have to actually provide that social proof.’

Often the proof in question doesn’t exist at all. Other times, it’s a plurality of ‘experts’ in a survey, or a reporter’s reading of a single scientific study, or three friends backing each other up – or people who have been told or gotten the impression everybody knows, so they claim to know, too.  The phrase ‘everybody knows’ is a great way to cause an information cascade.

The second central mode of ‘everyone knows’ is when it means ‘if you do not know this, or you question it, you are stupid, ignorant and blameworthy.’

It’s your own damn fault for going out in the rain and getting soaked. It’s your own damn fault for not knowing that everything politicians say (or something the speaker said) is a lie, even though they frequently tell the truth – which means they ‘aren’t really lies’ because no one was fooled. It’s your own damn fault for not keeping up with the latest gossip or fashion trends.

It is made clear that to question this is to show you are stupid, ignorant and blameworthy, especially if the statement everyone knows is false. You’d be all but volunteering to be the scapegoat.

A classic mode is the condemnation ‘everyone knows that X is (everywhere / great / the right thing / necessary / patriotic / fair / standard / appropriate / customary / the party line / how things get done around here / smart / right / a thing / not a thing / a conspiracy theory / wrong / evil / stupid / slander / rhetoric used by the out-group / rhetoric that supports the out-group / unacceptable / impossible / impractical / unthinkable / horrible / unfair / stupid / rude / your own fault / racist / sexist / treason / cheating / cultural appropriation / etc etc etc).

The whole point is to establish truth without allowing a response or providing evidence.

Note that this is self-referencing. To be someone, you have to know  what ‘everybody knows’ means.

A third central mode is ‘if you do not know this (and, often, also claim everyone knows this), you do not count as part of everyone, and therefore are no one. If you wish to be someone, or to avoid becoming no one, know this.’

This works both to make those on the outs not people, and to make the statements used unquestionable.

Thus, one is not blameworthy for acting as if everyone knows, because if someone is revealed not to know, that means they are no one, and therefore they have no

relevant impact or moral personhood. They can be ignored. Perhaps those who do not know this, or question it, are the outgroup. Perhaps they are simply those who don't get ahead, the little people. Perhaps they're just the fools we pity. Regardless, until they catch on, it is good and right to scam them – it is a sin to let a sucker keep his money.

A key variation on this is to flip the order into a way to admonish someone when they expose a falsehood or fraud someone wishes to perpetuate. First they argue that the thing is not a fraud, ideally that everyone knows it is not a fraud, but they lose, they fall back by flipping their position entirely. They now say: You're calling this thing a fraud. But *everyone knows* it's a fraud, so why are you wasting everyone's time saying it's a fraud when everyone already knows? This must be a social tactic, trying to lower the status of the fraud by pointing out what everyone already knows. Or if you think we don't already know, that must mean you think we aren't anyone. How insulting.

The fourth central mode is 'we are establishing this as true, and ideally as unquestionable, so pass that information along as something everyone knows.' It's aspirational, a self-fulfilling prophecy. Perhaps we already have done so by the time you're hearing this (and that's bad, because it means you're not hearing about new things everyone knows quickly enough!) or perhaps you're the first person to be told.

Either way, join the conspiracy. Spread that everybody knows the dice are loaded and rolls with their fingers crossed. Spread that everybody knows the war is over, and everybody knows the good guys lost.

So they'll cross their fingers rather than demand fair dice. So that they'll stop trying to fight the war.

# Does it become easier, or harder, for the world to coordinate around not building AGI as time goes on?

(Or, is coordination easier in a long timeline?)

It seems like it would be good if the world could coordinate to not build AGI. That is, at some point in the future, when some number of teams will have the technical ability to build and deploy and AGI, but they all agree to voluntarily delay (perhaps on penalty of sanctions) until they're confident that humanity knows how to align such a system.

Currently, this kind of coordination seems like a pretty implausible state of affairs. But I want to know if it seems like it becomes more or less plausible as time passes.

The following is my initial thinking in this area. I don't know the relative importance of the factors that I listed, and there's lots that I don't understand about each of them. I would be glad for…

- Additional relevant factors.
- Arguments that some factor is much more important than the others.
- Corrections, clarifications, or counterarguments to any of this.
- Other answers to the question, that ignore my thoughts entirely.

# If coordination gets harder overtime, that's probably because...

- **Compute increases make developing and/or running an AGI cheaper**. The most obvious consideration is that the cost of computing falls each year. If one of the bottlenecks for an AGI project is having large amounts of compute, then "having access to sufficient compute" is a gatekeeper criterion on who can build AGI. As the cost of computing continues to fall, more groups will be able to run AGI projects. The more people who can build an AGI, the harder it becomes to coordinate all of them into not deploying it.
    - Note that It is unclear to what degree there is currently, or will be, a hardware overhang. If someone in 2019 could already run an AGI, on only $10,000 worth of AWS, if only they knew how, then the cost of compute is not relevant to the question of coordination.
- **The number of relevant actors increases**. If someone builds an AGI in the next year, I am reasonably confident that that someone will be Deep Mind. I expect that in 15 years, if I knew that AGI would be developed one year from then, it will be much less overdetermined which group is going to build it, because there will be many more well funded AI teams with top talent, and, most likely, none of them will have as strong a lead as Deep Mind currently appears to have.
    - This consideration suggests that coordination gets harder over time. However, this depends heavily on other factors (like how accepted AI

safety memes are) that determine how easily Deep Mind could coordinate internally.

# If coordination gets easier over time, that's probably because…

- **AI safety memes become more and more pervasive and generally accepted.** It seems that coordination is easier in worlds where it is uncontroversial and common knowledge that an unaligned AGI poses and existential risk, because everyone agrees that they will lose big if anyone builds an AGI.
    - Over the past 15 years, the key arguments of AI safety have gone from being extremely fringe, to a reasonably regarded (if somewhat controversial) position, well inside the overton window. Will this process continue? Will it be commonly accepted by ML researches in 2030, that advanced AI poses and existential threat? Will it be commonly accepted by the leaders of nation-states?
    - What will the perception of safety be in a world where there is another AGI winter? Suppose that narrow ML proves to be extremely useful in a large number of fields, but there's lots of hype about AGI being right around the corner, then that bubble bursts, and there is broad disinterest in AGI again. What happens to the perception of AI safety? Is there a sense of "It looks like AI Alignment wasn't important after all"? How cautious will researchers be in developing new AI technologies.
- [Partial subpoint to the above consideration] **Individual AI teams develop more serious info security conscious processes.** If some team in Deep Mind discovered AGI today, and the Deep Mind leadership opted to wait to insure safety before deploying it, I don't know how long it would be until some relevant employees left to build AGI on their own, or some other group (such as a state actor) stole their technology and deployed it.
    - I don't know if this is getting better or worse, overtime.
- **The technologies for maintaining surveillance of would-be AGI developers improve**. Coordination is made easier by technologies that aid in enforcement. If surveillance technology improves that seems like it would make coordination easier. As a special case, highly reliable lie detection or mind reading technologies would be a game-changer for making coordination easier.
    - Is there a reason to think that offense will beat defense in this area? Surveillance could get harder over time if the technology for detecting and defeating surveillance outpaces the technology for surveilling.
- **Security technology improves**. Similarly, improvements in computer security (and traditional info security), would make it easier for actors to voluntarily delay deploying advanced AI technologies, because they could trust that their competitors (other companies and other nations), wouldn't be able to steal their work.
    - I don't know if this is plausible at all. My impression is that the weak point of all security systems is the people involved. What sort of advancements would make the human part of a security system more reliable?

# Learning biases and rewards simultaneously

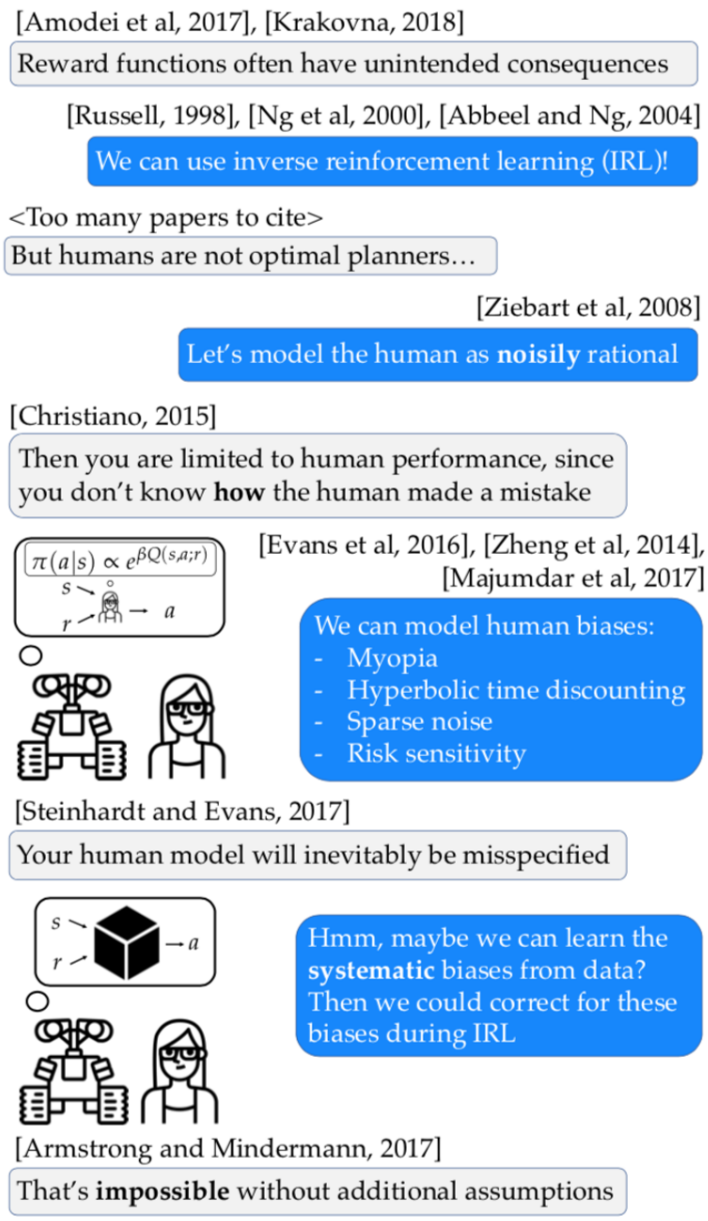
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I've finally uploaded to arXiv our work on inferring human biases alongside IRL, which was published at ICML 2019.

## Summary of the paper

### The IRL Debate

Here's a quick tour of the debate about inverse reinforcement learning (IRL) and cognitive biases, featuring many of the ideas from the first chapter of the Value Learning sequence:




I had the intuition that the impossibility theorem was like the other no-free-lunch theorems in ML: not actually relevant for what ML could do in practice. So we tried to learn and correct for systematic biases in IRL.

**The idea behind the algorithms**

The basic idea was to learn the *planning algorithm* by which the human produces demonstrations, and try to ensure that the planning algorithm captured the appropriate systematic biases. We used a Value Iteration Network to give an inductive bias towards "planners" but otherwise did not assume anything about the form of the systematic bias. [1] Then, we could perform IRL by figuring out which reward would cause the planning algorithm to output the given demonstrations. The reward would

be “debiased” because the effect of the biases on the policy would already be accounted for in the planning algorithm.

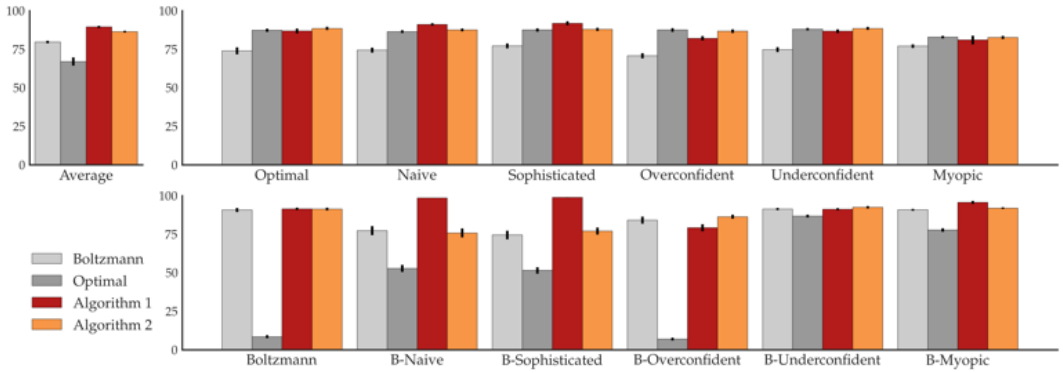
How could we learn the planning algorithm? Well, one baseline method is to assume that we have access to some tasks where the *rewards are known*, and use those tasks to learn what the planning algorithm is. Then, once that is learned, we can infer the rewards for new tasks that we haven’t seen before. This requires the planner to generalize across tasks.

However, it’s kind of cheating to assume access to ground truth rewards, since we usually wouldn’t have them. What if we learned the planning algorithm and rewards simultaneously? Well, the no-free-lunch theorem gets us then: maximizing the true reward and minimizing the negative of the true reward would lead to the same policy, and so you can’t distinguish between them, and so the output of your IRL algorithm could be the true reward or the *negative* of the true reward. It would be really bad if our IRL algorithm said exactly the opposite of what we want. But surely we can at least assume that humans are not expected utility *minimizers* in order to eliminate this possibility.

So, we make the assumption that the human is “near-optimal”. We initialize the planning algorithm to be optimal, and then optimize for a planning algorithm that is “near” the optimal planner, in gradient-descent-space, that combined with the (learned) reward function explains the demonstrations. You might think that a minimizer is in fact “near” a maximizer; empirically this didn’t turn out to be the case, but I don’t have a particularly compelling reason why that happened.

**Results**

Here’s the graph from our paper, showing the performance of various algorithms on some simulated human biases (higher = better). Both of our algorithms get access to the simulated human policies on multiple tasks. Algorithm 1 is the one that gets access to ground-truth rewards for some tasks, while Algorithm 2 is the one that instead tries to ensure that the learned planner is “near” the optimal planner. “Boltzmann” and “Optimal” mean that the algorithm assumes that the human is Boltzmann rational and optimal respectively.



Our algorithms work better on average, mostly by being robust to the specific kind of bias that the demonstrator had -- they tend to perform on par with the better of the Boltzmann and Optimal baseline algorithms. Surprisingly (to me), the second algorithm sometimes outperforms the first, even though the first algorithm has access to more data (since it gets access to the ground truth rewards in some tasks). This could be because it exploits the assumption that the demonstrator is near-optimal, which the first algorithm doesn't do, even though the assumption is correct for most of the models we test. On the other hand, maybe it's just random noise.

## Implications

### Superintelligent AI alignment

The most obvious way that this is relevant to AI alignment is that it is progress on ambitious value learning, where we try to learn a utility function that encodes all of human values.

"But wait,'' you say, "didn't you argue that ambitious value learning is unlikely to work?"

Well, yes. At the time that I was doing this work, I believed that ambitious value learning was the only option, and seemed hard but not doomed. This was the obvious thing to do to try and advance it. But this was over a year ago, the reason it's only now coming out is that it took a while to publish the paper. (In fact, it predates my state of the world work.) But it's true that *now* I'm not very hopeful about ambitious value learning, and so this paper's contribution towards it doesn't seem particularly valuable to me. However, a few others remain optimistic about ambitious value learning, and if they're right, this research might be useful for that pathway to aligned AI.

I do think that the paper contributes to narrow value learning, and I still think that this very plausibly will be relevant to AI alignment. It's a particularly direct attack on the specification problem, with the goal of inferring a specification that leads to a policy that would outperform the demonstrator. That said, I am no longer very optimistic about approaches that require a specific structure (in this case, world models fed into a differentiable planner with an inductive bias that then produces actions), and I am also less optimistic about using approaches that try to mimic expected value calculations, rather than trying to do something more like norm inference.

(However, I still expect that the impossibility result in preference learning will only be a problem in theory, not in practice. It's just that this particular method of dealing with it doesn't seem like it will work.)

### Near-term AI issues

In the near term, we will need better ways than reward functions to specify the behavior that we want to an AI system. Inverse reinforcement learning is probably the leading example of how we could do this. However, since the specific algorithms require much better differentiable planners before they will perform on par with existing algorithms, it may be some time before they are useful. In addition, it's probably better to use specific bias models in the near term. Overall, I think these methods or ideas are about as likely to be used in the near term as the average paper (which is to say, not very likely).

1. A Value Iteration Network is a fully differentiable neural network that embeds an approximate value iteration algorithm inside a feed-forward classification network. ↩

# The AI Timelines Scam

This is a linkpost for https://unstableontology.com/2019/07/11/the-ai-timelines-scam/

[epistemic status: that's just my opinion, man. I have highly suggestive evidence, not deductive proof, for a belief I sincerely hold]

"*If you see fraud and do not say fraud, you are a fraud.*" --- Nasim Taleb

I was talking with a colleague the other day about an AI organization that claims:

1. AGI is probably coming in the next 20 years.
2. Many of the reasons we have for believing this are secret.
3. They're secret because if we told people about those reasons, they'd learn things that would let them make an AGI even sooner than they would otherwise.

His response was (paraphrasing): "Wow, that's a really good lie! A lie that can't be disproven."

I found this response refreshing, because he *immediately* jumped to the most likely conclusion.

## Near predictions generate more funding

Generally, entrepreneurs who are optimistic about their project get more funding than ones who aren't. AI is no exception. For a recent example, see the Human Brain Project. The founder, Henry Makram, predicted in 2009 that the project would succeed in simulating a human brain by 2019, and the project was already widely considered a failure by 2013. (See his TED talk, at 14:22)

The Human Brain project got 1.3 *billion* Euros of funding from the EU.

It's not hard to see why this is. To justify receiving large amounts of money, the leader must make a claim that the project is actually worth that much. And, AI projects are more impactful if it is, in fact, possible to develop AI soon. So, there is an economic pressure towards inflating estimates of the chance AI will be developed soon.

## Fear of an AI gap

The missile gap was a lie by the US Air Force to justify building more nukes, by falsely claiming that the Soviet Union had more nukes than the US.

Similarly, there's historical precedent for an AI gap lie used to justify more AI development. Fifth Generation Computer Systems was an ambitious 1982 project by the Japanese government (funded for $400 million in 1992, or $730 million in 2019 dollars) to create artificial intelligence through massively parallel logic programming.

The project is widely considered to have failed. From a 1992 New York Times article:

> A bold 10-year effort by Japan to seize the lead in computer technology is fizzling to a close, having failed to meet many of its ambitious goals or to produce

> technology that Japan's computer industry wanted.
>
> ...
>
> That attitude is a sharp contrast to the project's inception, when it spread fear in the United States that the Japanese were going to leapfrog the American computer industry. In response, a group of American companies formed the Microelectronics and Computer Technology Corporation, a consortium in Austin, Tex., to cooperate on research. And the Defense Department, in part to meet the Japanese challenge, began a huge long-term program to develop intelligent systems, including tanks that could navigate on their own.
>
> ...
>
> **The Fifth Generation effort did not yield the breakthroughs to make machines truly intelligent, something that probably could never have realistically been expected anyway.** Yet the project did succeed in developing prototype computers that can perform some reasoning functions at high speeds, in part by employing up to 1,000 processors in parallel. The project also developed basic software to control and program such computers. Experts here said that some of these achievements were technically impressive.
>
> ...
>
> In his opening speech at the conference here, Kazuhiro Fuchi, the director of the Fifth Generation project, made an impassioned defense of his program.
>
> "Ten years ago we faced criticism of being too reckless," in setting too many ambitious goals, he said, adding, "Now we see criticism from inside and outside the country because we have failed to achieve such grand goals."
>
> Outsiders, he said, initially exaggerated the aims of the project, with the result that the program now seems to have fallen short of its goals.
>
> **Some American computer scientists say privately that some of their colleagues did perhaps overstate the scope and threat of the Fifth Generation project. Why? In order to coax more support from the United States Government for computer science research**.

(emphasis mine)

This bears similarity to some conversations on AI risk I've been party to in the past few years. The fear is that Others (DeepMind, China, whoever) will develop AGI soon, so We have to develop AGI first in order to make sure it's safe, because Others won't make sure it's safe and We will. Also, We have to discuss AGI strategy in private (and avoid public discussion), so Others don't get the wrong ideas. (Generally, these claims have little empirical/rational backing to them; they're based on scary stories, not historically validated threat models)

The claim that others will develop weapons and kill us with them by default implies a moral claim to resources, and a moral claim to be justified in making weapons in response. Such claims, if exaggerated, justify claiming more resources and making more weapons. And they weaken a community's actual ability to track and respond to real threats (as in The Boy Who Cried Wolf).

# How does the AI field treat its critics?

Hubert Dreyfus, probably the most famous historical AI critic, published "Alchemy and Artificial Intelligence" in 1965, which argued that the techniques popular at the time were insufficient for AGI. Subsequently, he was shunned by other AI researchers:

> The paper "caused an uproar", according to Pamela McCorduck. The AI community's response was derisive and personal. Seymour Papert dismissed one third of the paper as "gossip" and claimed that every quotation was deliberately taken out of context. Herbert A. Simon accused Dreyfus of playing "politics" so that he could attach the prestigious RAND name to his ideas. Simon said, "what I resent about this was the RAND name attached to that garbage."
>
> Dreyfus, who taught at MIT, remembers that his colleagues working in AI "dared not be seen having lunch with me." Joseph Weizenbaum, the author of ELIZA, felt his colleagues' treatment of Dreyfus was unprofessional and childish. Although he was an outspoken critic of Dreyfus' positions, he recalls "I became the only member of the AI community to be seen eating lunch with Dreyfus. And I deliberately made it plain that theirs was not the way to treat a human being."

This makes sense as anti-whistleblower activity: ostracizing, discrediting, or punishing people who break the conspiracy to the public. Does this still happen in the AI field today?

Gary Marcus is a more recent AI researcher and critic. In 2012, he wrote:

> Deep learning is important work, with immediate practical applications.
>
> ...
>
> Realistically, deep learning is only part of the larger challenge of building intelligent machines. Such techniques lack ways of representing causal relationships (such as between diseases and their symptoms), and are likely to face challenges in acquiring abstract ideas like "sibling" or "identical to." They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge, such as information about what objects are, what they are for, and how they are typically used. The most powerful A.I. systems ... use techniques like deep learning as just one element in a very complicated ensemble of techniques, ranging from the statistical technique of Bayesian inference to deductive reasoning.

In 2018, he tweeted an article in which Yoshua Bengio (a deep learning pioneer) seemed to agree with these previous opinions. This tweet received a number of mostly-critical replies. Here's one, by AI professor Zachary Lipton:

> There's a couple problems with this whole line of attack. 1) Saying it louder ≠ saying it first. You can't claim credit for differentiating between reasoning and pattern recognition. 2) Saying X doesn't solve Y is pretty easy. But where are your concrete solutions for Y?

The first criticism is essentially a claim that everybody knows that deep learning can't do reasoning. But, this is essentially admitting that Marcus is correct, while still criticizing him for saying it [ED NOTE: the phrasing of this sentence is off (Lipton

publicly agrees with Marcus on this point), and there is more context, see Lipton's reply].

The second is a claim that Marcus shouldn't criticize if he doesn't have a solution in hand. This policy deterministically results in the short AI timelines narrative being maintained: to criticize the current narrative, you must present your own solution, which constitutes another narrative for why AI might come soon.

Deep learning pioneer Yann LeCun's response is similar:

> Yoshua (and I, and others) have been saying this for a long time.
> The difference with you is that we are actually trying to do something about it, not criticize people who don't.

Again, the criticism is not that Marcus is wrong in saying deep learning can't do certain forms of reasoning, the criticism is that he isn't presenting an alternative solution. (Of course, the claim could be correct even if Marcus doesn't have an alternative!)

Apparently, it's considered *bad practice* in AI to criticize a proposal for making AGI without presenting on alternative solution. Clearly, such a policy causes large distortions!

Here's another response, by Steven Hansen (a research scientist at DeepMind):

> Ideally, you'd be saying this through NeurIPS submissions rather than New Yorker articles. A lot of the push-back you're getting right now is due to the perception that you haven't been using the appropriate channels to influence the field.

That is: to criticize the field, you should go through the field, not through the press. This is standard guild behavior. In the words of Adam Smith: "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices."

(Also see Marcus's medium article on the Twitter thread, and on the limitations of deep learning)

[ED NOTE: I'm not saying these critics on Twitter are publicly promoting short AI timelines narratives (in fact, some are promoting the opposite), I'm saying that the norms by which they criticize Marcus result in short AI timelines narratives being maintained.]

# Why model sociopolitical dynamics?

This post has focused on sociopolotical phenomena involved in the short AI timelines phenomenon. For this, I anticipate criticism along the lines of "why not just model the technical arguments, rather than the credibility of the people involved?" To which I pre-emptively reply:

- No one can model the technical arguments in isolation. Basic facts, such as the accuracy of technical papers on AI, or the filtering processes determining what you read and what you don't, depend on sociopolitical phenomena. This is far more true for people who don't themselves have AI expertise.

- "When AGI will be developed" isn't just a technical question. It depends on what people actually choose to do (and what groups of people actually succeed in accomplishing), not just what can be done in theory. And so basic questions like "how good is the epistemology of the AI field about AI timelines?" matter directly.
- The sociopolitical phenomena are actively making technical discussion harder. I've had a well-reputed person in the AI risk space discourage me from writing publicly about the technical arguments, on the basis that getting people to think through them might accelerate AI timelines (yes, really).

Which is not to say that modeling such technical arguments is not important for forecasting AGI. I certainly could have written a post evaluating such arguments, and I decided to write this post instead, in part because I don't have much to say on this issue that Gary Marcus hasn't already said. (Of course, I'd have written a substantially different post, or none at all, if I believed the technical arguments that AGI is likely to come soon had merit to them)

# What I'm not saying

I'm not saying:

1. That deep learning isn't a major AI advance.
2. That deep learning won't substantially change the world in the next 20 years (through narrow AI).
3. That I'm certain that AGI isn't coming in the next 20 years.
4. That AGI isn't existentially important on long timescales.
5. That it isn't possible that some AI researchers have asymmetric information indicating that AGI is coming in the next 20 years. (Unlikely, but possible)
6. That people who have technical expertise shouldn't be evaluating technical arguments on their merits.
7. That most of what's going on is people consciously lying. (Rather, covert deception hidden from conscious attention (e.g. motivated reasoning) is pervasive; see The Elephant in the Brain)
8. That many people aren't sincerely confused on the issue.

I'm saying that there are systematic sociopolitical phenomena that cause distortions in AI estimates, especially towards shorter timelines. I'm saying that people are being duped into believing a lie. And at the point where 73% of tech executives say they believe AGI will be developed in the next 10 years, it's a major one.

*This has happened before.* And, in all likelihood, *this will happen again*.

# Magic is Dead, Give me Attention

*(cross posted from my personal personal blog (not to be confused with my personal blog)*

## Back story

So here's the thing. I absolutely LOVE attention. I also HATE asking for attention. This past weekend I've processed those last two statements on a much deeper level than I have previously. Sometimes you need to rediscover an insight multiple times to really get it.

The most recent batch of introspection was prompted by the book "Magic is Dead" by Ian Frisch. Spoiler, I'm a semi-professional close up magician and have a vested interest in many things magic. Ian was describing Chris Ramsay, the quintessential young-blood social media based cool kid magician. A quote:

> [...] Ramsay's style has since evolved into a more high-end street-wear, hypebeast-esque aesthetic: A Bathing Ape jacket, Supreme cap and hoodie, adidas by Pharrell Williams NMD sneakers, etc.

I don't even know what A Bathing Ape jacket looks like, nor why the "A" is capitalized, but my gut reaction is revulsion. I know just enough about Supreme and "hypebeast"s to know that I don't like them. But why? Why do I spit venom when I hear about new performers on social media trying to "make magic cool again"?

Long story short, I LOVE attention, and I HATE asking for it. This has been the case since middle school. The hating to ask part comes with an attitude of "Fuck you, I'm not going to beg for you to give me something". So I had a chip on my shoulder in

regards to asking for people's attention, and like any good chip, it needed a rationalized narrative to justify its existence. "People who need other people to pay attention to them are weak", "Wow hazard, you're such a strong rugged individual for not needing other people's attention"

Notice the switch from "I don't like asking for attention" to "I don't need attention". It's subtle. Sure took me 8+ years to notice.

As you may have guessed, I didn't magically stop wanting/needing attention from people in middle school. I just learned a new strategy to meet the need. The strategy was the personality that I began to develop in middle school and high school, that of the unflappable competent marauder. I always played calm and collected, I worked hard to get good at the tasks at hand (my primary social group was my boy scout troop, so this meant getting good at outdoorsing, leadership, and planning), and casually leveraging my more impressive abilities (I didn't do parkour back then, but I could still climb most things and do dive rolls off of pavilion roof tops).

I become the sort of person who when you find out they also know how to juggle you go, "Of *course* Hazard knows how to juggle!" (actual interaction).

When I got into magic, tI worked this angle on a whole new level. Now I had a skill set where I could blow peoples minds and setup scenarios where *of course* we're all now going to watch Hazard do a magic trick, because they're so god damn cool!

Did I ever mention that I love attention? But really, in a non snarky, non self deprecating way, I love attention. Having people laughing and smiling and shouting with me at the center is such a yummy experience. A++, would recommend.

So my implicit approach to magic (and all relationships) was this: casually be as awesome as possible so that people are **compelled** to give me their attention, that way **I don't have to ask for it**.

# A "detour" into thanking people

Here's an implicit model of thanks and appreciation I see people using: If someone goes out of their way to help you, or wasn't expected to help you in the first place, thank them more. If it was easy for someone to help you, or it was expected of them, thank them less.

The counter to this would be telling you "don't take people for granted". This phrase always felt a bit odd to me. I'm not allowed to take anything for granted? Do I need to thank the strangers I passed on the way to this coffee shop because the didn't try to kill me? Seems a little much.

Here's a new framing: In the first model, one uses thanks and appreciation as a **marker of social debt**. If something happened such that "I owe you one" (small or big), then I mark it with a "thanks man". [Ignores for now that some people also seem to thank and appreciate as needed to make people do stuff for them]

The nugget of gold that I see in "don't take people for granted" is "let people know when they've helped you out". People want to be effectual, and it's a nice little boost to know that something you did actually helped someone. Because of how common the"thanks as debt marker" mentality is ("make sure you thank the neighbors for the extra dessert they let you have, they are very nice people and didn't have to give you

that") I think defenders of "granted" get roped into using the language of debt, thus leading to claims like "You owe everybody everything".

So there are two separate questions. When do you **owe someone something**, and when should you **thank or appreciate them**?

Owing is a huge beast on its own. There's a whole host of hidden sub questions. How should I personally treat people? When should I feel obligated to help someone? When should I be socially held accountable for helping someone? Complex stuff, not the topic of this post.

# Effects on the personal

Two things are extra hard if you hate to ask for attention. It's hard to **show appreciation**, and it's hard to **ask for things**.

Me trying to compel people to want to be around me instead of making explicit bids for friendship was a way of protecting myself from feeling like I owed anyone. "You're not doing me a favor by hanging out with me, because I'm so shiny you can't *not* hang out with me." And then, oh oops, this leads to rarely showing appreciation. Now, remember that I'm not asking people for stuff all the time. The failure mode here is not "I'm always doing all this stuff for you and you never appreciate it!". It's a bit more subtle.

(paraphrased quote someone has said to me)

> Yeah, you're cool and all but I don't know why you hang out with me. I mean it doesn't seem like you get much out of it. It doesn't feel like I could actually matter that much to you."

Yikes.

Also notice the lovely way this can be more extreme if a friend has low self esteem (which may or may not be a factor that makes someone more likely to be compelled to a shiny person(?)).

Yikes.

So yeah, not showing people appreciation ain't cool.

Not asking for things, that's more of a me problem. "Good things come to those who ask" and all that.

# Show Biz

Zooming back to show biz and performance. Recall, my approach with magic was be so good that people were compelled to give me their attention, and rarely if ever make explicit bids for attention. If you keep scaling this attitude up, you roughly get "Fuck you, I'm awesome. Come see my show if you want to have a good time, it's your loss if you don't." There are some important things that this frame gets right.

If you as a performer feel destroyed every time someone doesn't like your show, or when the theater isn't booked solid, you're in for a world of hurt. It's also likely you will

have a hard time developing your own style. If you're terrified of being disliked, you face huge pressure to play it safe and stick to the known. A certain about of "Fuck you, I'm awesome" is needed to be yourself. How much of it is needed? Hard to say.

Second point, people need a reason why they should be watching *you* as opposed to the millions of other options they have. People want to see stuff they are going to enjoy, and if you don't at least say, "Yes, my show is in fact good and you will enjoy it" lots of people will just move onto something else.

Those are the good parts, now here's the poison in it. "Fuck you, I'm awesome" is a frame that asserts that you, the audience, don't matter. Your attention/time/money most not matter if I don't care if you come to my show.

No, I'm not saying that every performer tell their fans they care about them. You might not care about your fans, you might not even know them. But I do want to explore what it would be like to both give incredible performances that people love and enjoy, while also explicitly appreciating the good thing we have with and letting them know they are doing a good thing for me.

# "Can I get this to go please?"

What if you, whenever people did good stuff for you, you let them know they had a positive effect on you? You might have to come up with a unique way of saying it if you want to make explicit that you aren't saying "I owe you". We'll leave that as an exercise to the reader.

# What are we predicting for Neuralink event?

Interesting exercise in AI-adjacent forecasting area (brain-computer interfaces). Curious if people want to specify some possible reveals+probabilities.
https://twitter.com/neuralink/status/1149133717048188929

(if in the somewhat likely scenario you're relying on inside info please mention it)

# How much background technical knowledge do LW readers have?

When writing posts, it would be useful to know how much background technical knowledge LW readers have in various areas.

To that end, I set up a short six-question survey. Please take it, so I (and others) can write posts better fit to your level of technical background. If your answer to all of the questions is "zero-ish technical knowledge", please take it, so you're not inundated with mathy posts. If your answer to all of the questions is "I am secretly John Von Neumann", please take the survey, so the rest of us know there's someone like that around. If you are somewhere in the middle, please take the survey. It'll take, like, maybe sixty seconds.

Here's that link again: survey.

# Book review: The Technology Trap

I recently finished reading *The Technology Trap*, by Carl Frey. The book attempts to do two things: chronicle the role of technology in economic progress throughout history, and argue that automation in our own era parallels the first seven decades of the industrial revolution, during which the wealth from mechanisation failed to reach most of the population.

I particularly enjoyed the first component, because until now I've read much less about the industrial revolution than I should have. That also means that I'm not qualified to evaluate the book's accuracy. However, it had interesting discussions of:

- The technological prowess of the Romans, and why they were held back from industrialising both because of their slave-based economy, and also because of an implicit dismissal of the private economy.
- The development of some surprisingly important technologies during the Middle Ages, such as wind- and water-mills, better ways of accessing horsepower (via improved horseshoes, harnesses and ploughs), and town clocks.
- The fact that most of the key innovations of the early industrial revolution (steam engines excepted) would have been technologically possible a century or two earlier, but were blocked by the political power of guilds.
- The importance of the Glorious Revolution in shifting England's political climate to favour industrialisation; and more generally, the role of competition between nation-states in spurring government permissiveness towards innovation.
- The mechanisation of the silk industry as a smaller-scale precursor to the mechanisation of cotton processing that would drive the early years of the industrial revolution.
- The prevalence of child labour in the first factories run by Arkwright and others; and more generally how miserable the first few decades of the industrial revolution were for the poor, who were crammed into unsanitary cities on reduced wages, with severe health consequences.
- The fact that it took many decades after Watt's steam engine was patented in 1769 for railways to actually become widely significant.
- The role of arms manufacturers like Colt as precursors to Ford's assembly lines.

In its latter role, however, the book seems a little incomplete. From around 1770 to 1840, productivity rose while worker incomes stagnated, with the increased wealth primarily going to industrialists - a period known as "Engels' pause". Frey argues that today, as the incomes of Western workers stagnate, we've reached an analogous situation. Engels' pause gave rise to Luddite riots and the growth of the communist movement. Similarly, the modern working class will be tempted to campaign against automation - a "technology trap" which we will need to overcome to reach the level of technology which makes prosperity more widespread.

Certainly the thesis is initially plausible, but at the end of the book I was left with quite a few unanswered questions. Four particularly important ones:

1. Frey makes the distinction between replacing technologies and augmenting technologies. The former "render jobs and skills redundant"; automatic elevators are a good example. The latter "make people more productive in existing tasks or create entirely new jobs for them"; Frey's examples are innovation in the steel

industry and the invention of the typewriter. But there's a pretty blurry line between these two categories. An augmenting technology becomes a replacing technology if "demand for a given product or service becomes saturated", a criterion which has less to do with the sector itself than with the broader state of the economy. But if we're considering the wider economy, then the lower costs provided by replacing technologies enable other sectors to produce more goods, making them augmenting after all. So while the call to "augment not replace" workers has become a rallying cry, I'm not sure that the distinction has much predictive power. Can we tell in advance which technologies will be augmenting vs replacing, or do we just have to wait until a few decades later and look at the job statistics?

2. Building on the last point: you could describe the first industrial revolution as starting off with replacing technologies (such as power looms) and moving on to augmenting technologies (such as the steam engine). And you could describe the second industrial revolution as being all about augmenting technologies (such as electricity and cars - although the latter could also be considered a replacing technology for horses). If Frey is right that the current wage stagnation has been driven by automation, then this matches the beginning of the first industrial revolution.* But are there good reasons to think that we'll eventually transition to building augmenting technologies in the same way as they did? Reasoning from a small sample size is treacherous at the best of times, and in this case our n=2 sample showcases two different trajectories. We might be about to experience a third distinct trajectory: AI continuing to be a replacing technology to a greater and greater extent. I do think this is unlikely (as I argue here) but it's an open possibility.
3. Frey discusses the experience of America's blue-collar middle class - which, he argues, has lost jobs to a combination of globalisation and automation. But (assuming this is true) how much of the responsibility should each factor bear? If it's almost all due to globalisation, then the chapter is a little misleading. I don't have any particular reason to think that, but Frey doesn't do the work of convincing me otherwise. (Although, since globalisation has been made much easier by information technology, should we count it as an effect of automation? It seems roughly analogous to how technologies invented early in the industrial revolution allowed adults' jobs to be done by children.)
4. Frey worries that the technology trap will lead to workers suppressing technological growth. Yet there have been many changes to the factors which originally held back industrialisation. Guilds/unions are much reduced in power; international competitiveness is now a top priority; faster communication channels facilitate the spread of new ideas; and the intellectual plausibility of stifling innovation as a way to protect workers is much diminished, given how hugely we have benefited (in material terms) from the last few centuries of technological progress.
   On the other hand, everyone has the vote now, which wasn't the case in the past. And many people are using those votes to send a strong message against current intellectual orthodoxy. And with the pace of change being much faster now than in the 1700s, perhaps the backlash it spurs will be concomitantly greater. Or maybe it will mean that the anti-technology camp has less time to coordinate resistance. It seems very unclear how these factors weigh against each other; Frey's historical analogy can only take us so far.

Frey finishes with a set of prescriptions for how to close the gap between the winners and losers from automation, most of which are standard and sensible - e.g. cutting back on occupational licensing, encouraging relocation, investing in high-speed rail,

and reforming housing markets. A more novel proposal is wage insurance, which compensates people when they are forced into lower-paying jobs. It seems like a good idea for individuals, but if implemented by the government as Frey suggests, I worry that it'll become yet another piece of clutter in an already overcomplicated and inefficient welfare system.

I want to end this review with a theme of the book that I particularly liked: the rehabilitation of the Luddites. Frey emphasises that, despite having become a byword for ignorant destructiveness, the Luddites were actually campaigning against a major threat to their livelihoods and communities, and we should sympathise with them. The parallels with our modern era are obvious - and the more we can rise above pejorative descriptions of our political opponents, the better.

* There's also the complication that incomes in the tech sector have been rising rapidly. Was there an analogous group of skilled workers who benefited from Engels' pause? I suppose that the job of building the machines must have been a lucrative one, but I really don't know.

# Wolf's Dice

Around the mid-19th century, Swiss astronomer Rudolf Wolf rolled a pair of dice 20000 times, recording each outcome. Here is his data:

| | | White Die | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Red Die | 1 | 547 | 587 | 500 | 462 | 621 | 690 | 3407 |
| | 2 | 609 | 655 | 497 | 535 | 651 | 684 | 3631 |
| | 3 | 514 | 540 | 468 | 438 | 587 | 629 | 3176 |
| | 4 | 462 | 507 | 414 | 413 | 509 | 611 | 2916 |
| | 5 | 551 | 562 | 499 | 506 | 658 | 672 | 3448 |
| | 6 | 563 | 598 | 519 | 487 | 609 | 646 | 3422 |
| | Total | 3246 | 3449 | 2897 | 2841 | 3635 | 3932 | 20000 |

I've taken the data from Jaynes, who uses it as an example for maximum-entropy methods. We'll use it for several examples, including some of Jaynes' models.

The most interesting fact about Wolf's dice data is that some faces come up significantly more often than others. As a quick back-of-the-envelope check, we can see this by calculating the expected number of times a face should come up on one die ($20000 * \frac{1}{6} \approx 3333.3$), and the standard deviation of this number ($\sqrt{20000 * (\frac{1}{6}) * (1 - \frac{1}{6})} \approx 52.7$). A quick glance at the data shows that the column- and row-totals differ from their expected value by roughly 2 to 6 standard deviations, an error much larger than we'd expect based on random noise.

That, however, is an ad-hoc test. The point of this sequence is to test our hypotheses from first principles, specifically the principles of probability theory. If we want to know which of two or more models is correct, then we calculate the probability of each model given the data.

# White Die: Biased or Not?

Let's start with just Wolf's white die. Our two models are:

- Model 1: all faces equally probable
- Model 2: each face i has its own probability $p_i$ (we'll use a uniform prior on $p_i$ for now, to keep things simple)

Our data is the bottom row in the table above, and our goal is to calculate $P[\text{model}_i|\text{data}]$ for each model.

The first step is Bayes' rule:

$$P[\text{model}_i|\text{data}] = \frac{1}{Z}P[\text{data}|\text{model}_i]P[\text{model}_i]$$

Here's what those pieces each mean:

- $P[\text{model}_i]$ is our prior probability for each model - for simplicity, let's say the two are equally likely a priori.

- Z is the normalizer, chosen so that the posterior probabilities sum to one: $P[\text{model}_1|\text{data}] + P[\text{model}_2|\text{data}] = 1$ (implicitly, we assume one of the two models is "correct").
- $P[\text{data}|\text{model}_i]$ is computed using $\text{model}_i$

... so the actual work is in calculating $P[\text{data}|\text{model}_i]$.

# Model 1: Unbiased

For the first model, $P[\text{data}|\text{model}_1]$ is a standard probability problem: given $n = 20000$ unbiased die rolls, what's the probability of $n_1 = 3246$ 1's, $n_2 = 3449$ 2's, etc? This is a multinomial distribution; the answer is

$$P[n_1 \dots n_6|\text{model}_1] = \frac{n!}{n_1! \dots n_6!} p_1^{n_1} \dots p_6^{n_6} = \frac{20000!}{3246! \dots 3312!} \left(\frac{1}{6}\right)^{20000} \approx 6.3 * 10^{-70}$$

Note the symmetry factor with the factorials: we're computing the probability of the observed *counts*, not the probability of a particular string of outcomes, so we have to add up probabilities of all the outcomes with the same counts. The same factor will appear in model 2 as well, for the same reason.

# Model 2: Biased

The second model is a bit more complicated. If we knew the probabilities for each face $p_1 \dots p_6$ then we could use the multinomial distribution formula, same as model 1. But since we don't know the p's, we need to integrate over possible values:

$$P[\text{data}|\text{model}_2] = \int_p P[\text{data}|p] dP[p] = \int_p \frac{n!}{n_1! \dots n_6!} p_1^{n_1} \dots p_6^{n_6} dP[p]$$

Here $dP[p]$ is our prior distribution on p - in this case uniform, i.e. a dirichlet distribution with parameter $\alpha = 1$. This is a somewhat nontrivial integral: the constraint $\sum_i p_i = 1$ means that we're integrating over a five-dimensional surface in six dimensions. Fortunately, other people have already solved this integral in general: it's the very handy dirichlet-multinomial distribution. With $\alpha = 1$ the result is particularly simple; the integral comes out to:

$$\frac{n!}{n_1! \dots n_k!} \int_p p_1^{n_1} \dots p_k^{n_k} dP[p|\alpha = 1] = \frac{n!(k-1)!}{(n+k-1)!}$$

We'll use this formula quite a bit in the next post. For now, we're using $k = 6$ outcomes, so we get

$P[data|model_2] = \int_p P[data|p]dP[p] = \frac{20000!}{20005!} \approx 3.7 * 10^{-20}$

So the data is around $10^{50}$ times more likely under this model than under the unbiased-die model.

# Probability of Bias

Last step: let's go back to where we started and compute the posterior probabilities of each model. We plug $P[data|model_i]$ back into Bayes' rule. Each model had a prior probability of 0.5, so we compute the normalizer Z as:

$Z = P[data|model_1]P[model_1] + P[data|model_2]P[model_2] \approx 1.9 * 10^{-20}$

so

$P[model_1|data] = \frac{1}{Z} P[data|model_1]P[model_1] \approx 1.7 * 10^{-50}$

$P[model_2|data] \approx 1.0 - 1.7 * 10^{-50}$

A few comments on this result...

First, we have pretty conclusively confirmed that the faces are not equally probable, given this data.

Second, the numbers involved are REALLY BIG - ratios on the order of $10^{50}$. This is par for the course: since independent probabilities multiply, probabilities tend to be roughly exponential in the number of data points. One side effect is that, as long as we have a reasonable amount of data, the priors don't matter much. Even if we'd thought that an unbiased die was a thousand times more likely than a biased die a priori, those huge exponents would have completely swamped our prior.

Playing around with α will reveal that the same applies to our prior distribution for p: the result is not very sensitive to changes in the prior. A word of caution, however: priors over unknown parameters become more important in high-dimensional problems.

Third, notice that there was no maximization and no extra parameters to twiddle. Once the models were specified, we had zero choices to make, zero degrees of freedom to play with. Any "free parameters" - i.e. p - have a prior over which we integrate. That integral is the hard part: as the models get larger and more complicated, we need to evaluate hairy high-dimensional integrals. The problem is not just NP-complete, it's #P-complete. In practice, approximations are used instead, including the entire range of hypothesis tests and model comparison tests - that's where maximization enters the picture. We'll talk about some of those later, especially about when they're likely to work or not work.

Finally, note that we compared a model which is conceptually "compatible" with the data (i.e. capable of generating roughly the observed frequencies), to one which is conceptually "incompatible" (i.e. the observed frequencies are way outside the expected range). A more interesting case is to consider two models which are both "compatible". In that case, we'd want to use some kind of complexity penalty to say that the more complex model is less

likely - otherwise we'd expect overfit. In the next post, we'll revisit Wolf's dice with a couple models from Jaynes, and see how P[data|model] "penalizes" overly-complicated models.

# What does Optimization Mean, Again? (Optimizing and Goodhart Effects - Clarifying Thoughts, Part 2)

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

**Clarifying Thoughts on Optimizing and Goodhart Effects - Part 2**

Previous Post: Re-introducing Selection vs Control for Optimization In the post, I reviewed Abram's selection/control distinction, and suggested how it relates to actual design. I then argue that there is a bit of a continuum between the two cases, and that we should add an addition extreme case to the typology, direct solution.

Here, I will revisit the question of what optimization means.

NOTE: This is not completely new content, and is instead split off from the previous version and rewritten to include an *(Added)* discussion of Eliezer's definition for measuring optimization power, from 2008. Hopefully this will make the sequence clearer for future readers.

In the next post, Applying over-Optimization in Selection and Control, I apply these ideas, and concretize the discussion a bit more before moving on to discussing Mesa-Optimizers in Part 4.

# What does Optimization Mean, Again?

This question has been discussed a bit, but I still don't think its clear. So I want to start by revisiting a post Eliezer wrote in 2008, where he suggested that optimization power was ability to select states from a preference ordering over different states, and could be measured with entropy. He notes that this is not computable, but gives us insight. I agree, except that I think that the notion of the state space is difficult, for some of the reasons Scott discussed when he mentioned that he was confused about the relationship between gradient descent and Goodhart's law. In doing so, Scott proposed a naive model that looks very similar to Eliezer's;

> simple proxy of "sample points until I get one with a large U value" or "sample n points, and [*select*] the one with the largest U value" when I think about what it means to optimize something for U. I might even say something like "n bits of optimization" to refer to sampling $2^n$ points. I think this is not a very good proxy for what most forms of optimization look like."

I want to start by noting that this is absolutely and completely a "selection" type of optimization, in Abram's terms. As Scott noted, however, it's not a good model for what most optimization looks like, and that's part of why I think Eliezer's model is less helpful than I did when I originally read it.

There's a much better model for gradient descent optimization, which is... gradient descent. It is a bit closer to control than direct optimization, since in some sense we're navigating through the space, but for almost all actual applications, it is still selection, not control. To review how it works, points are chosen iteratively, and the gradient is assessed at each point. The gradient is used to *select* a new point at some (perhaps very clever, dynamically chosen next point.) Some stopping criteria is checked, and it iterates at that new point. This is almost always tons more efficient than generating random points and examining them.

*(Addded)* It's far better than a grid search, usually, for most landscapes, but also makes it clear why I think it's hard to discuss optimization power in Eliezer's terms on a practical level, at least when dealing with a continuous system. The problem I'm alluding to is that any list of preferences over states depends on number of states. Gradient descent type optimization is really good at focusing on specific sections of the state space, especially compared to grid search. We might find a state where grid search would require a tremendously high resolution, but we don't ever compute a preference ordering over $2^n$ states. With gradient descent, we instead compute preferences for a local area and (hopefully) zoom-in, potentially ignoring other parts of the space. An optimizer that focuses very narrowly can have high-resolution but miss the non-adjacent region with far better outcomes, or can have fairly low resolution but perform far better - and the second optimizer is clearly more powerful, but I don't know how to capture this.

But to return to the main discussion, the process of gradient descent is also somewhere between selection and control - and that's what I want to explain.

In theory, the evaluation of each point in the test space could involve an actual check of the system. I build each rocket, watch to see whether it fails or succeeds according to my metric. For search, I'd just pick the best performers, and for more clever approaches, I can do something like find a gradient by judging performance of parameters to see if increasing or decreasing those that are amenable to improvement would help. (I can be even more inefficient, but find something more like a gradient, by building many similar rockets, each an epsilon away in several dimensions, and estimating a gradient that way. Shudder.)

In practice, we use a proxy model - and this is one place that allows for the types of overoptimization misalignment we are discussing. (But it's not the only one.) The reason this occurs is laid out clearly in the Categorizing Goodhart paper as one of the two classes of extremal failure - either model insufficiency, or regime change. This also allows for (during simulation undetectable) causal failures, if the proxy model gets a causal effect wrong.Even without using a proxy model, we can be led astray by the results if we are not careful. Rockets might look great, even in practice, and only fail in untested scenarios because we optimized something too hard - extremal model insufficiency. (Lower weight is cheaper, and we didn't notice a specific structural weakness induced by ruthlessly eliminating weight on the structure.) For our purposes, we want to talk about things like "how much optimization pressure is being applied." This is difficult, and I think we're trying to fit incompatible conceptual models together rather than finding a good synthesis, but I have a few ideas on what selection pressure leading to extremal regions means here.

- Extreme proxy values (in comparison to most of the space) seems similar to having lots of selection pressure. If we have a insanely tall and narrow peak, we may be finding something strange rather than simply improving.

- Extreme input values (unboundedly large or small values) may indicate a worrying area vis-a-vis overoptimization failures.
- Lots of search time alone does NOT indicate extremal results - it indicates lots of things about your domain, and perhaps the inefficiency of your search, but not overoptimization. (This is in contrast to the naive grid-search model, where lots of points visited means more optimizing.)

As an aside, Causal Goodhart is different. It doesn't really seem to rely on extremes, but rather on manipulating new variables, ones that could have an impact on our goal. This can happen because we change the value to a point where it changes the system, similar to extremal Goodhart, but does not need to. For instance, we might optimize filling a cup by getting the water level near the top. Extremal regime change failure might be overfilling the cup and having water spill everywhere. Causal failure might be moving the cup to a different point, say right next to a wall, in order to capture more water, but accidentally break the cup against the wall.Notice that this doesn't require much optimization pressure - Causal Goodhart is about moving to a new region of the distribution of outcomes by (metaphorically or literally) breaking something in the causal structure, rather than by over-optimizing and pushing far from the points that have been explored.This completes the discussion so far - and note that none of this is about control systems. That's because in a sense, most current examples don't optimize much, they simply execute an adaptive program.

One critical case of a control system optimizing is a mesa-optimizer, but that will be deferred until after the next post, which introduces some examples and intuitions around how Goodhart-failures occur in selection versus control systems.

# June 2019 gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2019/06

# Why artificial optimism?

This is a linkpost for https://unstableontology.com/2019/07/15/why-artificial-optimism/

Optimism bias is well-known. Here are some examples.

- It's conventional to answer the question "How are you doing?" with "well", regardless of how you're actually doing. Why?
- People often believe that it's inherently good to be happy, rather than thinking that their happiness level should track the actual state of affairs (and thus be a useful tool for emotional processing and communication). Why?
- People often think their project has an unrealistically high chance of succeeding. Why?
- People often avoid looking at horrible things clearly. Why?
- People often want to suppress criticism but less often want to suppress praise; in general, they hold criticism to a higher standard than praise. Why?

# The parable of the gullible king

Imagine a kingdom ruled by a gullible king. The king gets reports from different regions of the kingdom (managed by different vassals). These reports detail how things are going in these different regions, including particular events, and an overall summary of how well things are going. He is quite gullible, so he usually believes these reports, although not if they're too outlandish.

When he thinks things are going well in some region of the kingdom, he gives the vassal more resources, expands the region controlled by the vassal, encourages others to copy the practices of that region, and so on. When he thinks things are going poorly in some region of the kingdom (in a long-term way, not as a temporary crisis), he gives the vassal fewer resources, contracts the region controlled by the vassal, encourages others not to copy the practices of that region, possibly replaces the vassal, and so on. This behavior makes sense if he's assuming he's getting reliable information: it's better for practices that result in better outcomes to get copied, and for places with higher economic growth rates to get more resources.

Initially, this works well, and good practices are adopted throughout the kingdom. But, some vassals get the idea of exaggerating how well things are going in their own region, while denigrating other regions. This results in their own region getting more territory and resources, and their practices being adopted elsewhere.

Soon, these distortions become ubiquitous, as the king (unwittingly) encourages everyone to adopt them, due to the apparent success of the regions distorting information this way. At this point, the vassals face a problem: while they want to exaggerate their own region and denigrate others, they don't want others to denigrate their own region. So, they start forming alliances with each other. Vassals that ally with each other promise to say only good things about each other's regions. That way, both vassals mutually benefit, as they both get more resources, expansion, etc compared to if they had been denigrating each other's regions. These alliances also make sure to keep denigrating those not in the same coalition.

While these "praise coalitions" are locally positive-sum, they're globally zero-sum: any gains that come from them (such as resources and territory) are taken from other regions. (However, having more praise overall helps the vassals *currently in power*, as it means they're less likely to get replaced with other vassals).

Since praise coalitions lie, they also suppress *the truth in general* in a coordinated fashion. It's considered impolite to reveal certain forms of information that could imply that things aren't actually going as well as they're saying it's going. Prying too closely into a region's actual state of affairs (and, especially, sharing this information) is considered a violation of privacy.

Meanwhile, the actual state of affairs has gotten worse in almost all regions, though the regions prop up their lies with Potemkin villages, so the gullible king isn't shocked when he visits the region.

At some point, a *single* praise coalition wins. Vassals notice that it's in their interest to join this coalition, since (as mentioned before) it's in the interests of the vassals *as a class* to have more praise overall, since that means they're less likely to get replaced. (Of course, it's also in their class interests to have things *actually* be going well in their regions, so the praise doesn't get too out of hand, and criticism is sometimes accepted) At this point, it's conventional for vassals to always praise each other and punish vassals who denigrate other regions.

Optimism isn't ubiquitous, however. There are a few strategies vassals can use to use pessimism to claim more resources. Among these are:

- *Blame*: By claiming a vassal is doing something wrong, another vassal may be able to take power away from that vassal, sometimes getting a share of that power for themselves. (Blame is often not especially difficult, given that everyone's inflating their impressions)
- *Pity*: By showing that their region is undergoing a temporary but fixable crisis (perhaps with the help of other vassals), vassals can claim that they should be getting more resources. But, the problem has to be solvable; it has to be a temporary crises, not a permanent state of decay. (One form of pity is claiming to be victimized by another vassal; this mixes blame and pity)
- *Doomsaying*: By claiming that there is some threat to the kingdom (such as wolves), vassals can claim that they should be getting resources in order to fight this threat. Again, the threat has to be solvable; the king has little reason to give someone more resources if there is, indeed, nothing to do about the threat.

Pity and doomsaying could be seen as two sides of the same coin: pity claims things are going poorly (but fixably) locally, while doomsaying claims things are going poorly (but fixably) globally. However, all of these strategies are limited to a significant degree by the overall praise coalition, so they don't get out of hand.

# Back to the real world

Let's relate the parable of the gullible king back to the real world.

- The king is sometimes an actual person (such as a CEO, as in Moral Mazes, or a philanthropist), but is more often a process distributed among many people that is evaluating which things are good/bad, in a pattern-matching way.

- Everyone's a vassal to some degree. People who have more power-through-appearing-good are vassals with more territory, who have more of an interest in maintaining positive impressions.
- Most (almost all?) coalitions in the real world have aspects of praise coalitions. They'll praise those in the coalition while denigrating those outside it.
- Politeness and privacy are, in fact, largely about maintaining impressions (especially positive impressions) through coordinating against the revelation of truth.
- Maintaining us-vs-them boundaries is characteristic of the political right, while dissolving them (and punishing those trying to set them up) is characteristic of the political left. So, non-totalizing praise coalitions are more characteristic of the right, and total ones that try to assimilate others (such as the one that won in the parable) are more characteristic of the left. (Note, totalizing praise coalitions still denigrate/attack ones that can't be safely assimilated; see the paradox of tolerance)
- Coalitions may be fractal, of course.
- A lot of the distortionary dynamics are subconscious (see: The Elephant in the Brain).

This model raises an important question (with implications for the real world): if you're a detective in the kingdom of the gullible king who is at least somewhat aware of the reality of the situation and the distortonary dynamics, and you want to fix the situation (or at least reduce harm), what are your options?

# Doublecrux is for Building Products

Previously:

- What product are you building?

2 years ago, CFAR's doublecrux technique seemed "probably good" to me, but I hadn't really stress tested it. And it was particularly hard to learn in isolation without a "real" disagreement to work on.

Meanwhile, some people seemed skeptical about it, and I wasn't sure what to say to them other than "I dunno man this just seems obviously good? Of *course* you want to treat disagreements like an opportunity to find truth together, share models, and look for empirical tests you can run?"

But for people who didn't share that "of course", that wasn't very helpful.

For the past two years I've worked on a team where big disagreements come up pretty frequently, and where doublecrux has been more demonstrably helpful. I have a clearer sense of where and when the technique is important.

# Intractable Disagreements

**Some intractable disagreements are fine**

If you disagree with someone on the internet, or a random coworker or something, often the disagreement doesn't matter. You and your colleague will go about their lives, one way or another. If you and your friends are fighting over "Who would win, Batman or Superman?", coming to a clear resolution just isn't the point.

It might also be that you and your colleague are doing some sort of coalition-politics fight over the overton window, and most of the debate might be for the purpose of influencing the public. Or, you might be arguing about the Blue Tribe vs Red tribe as a way of signaling group affiliation, and earnestly understanding people isn't the point.

This makes me sad, but I think it's understandable and sometimes it's even actually important.

Such conversations are don't *need* to be doublecrux shaped, unless both participants want them to be.

**Some disagreements are *not* fine**

When you're building a product together, it actually matters that you figure out how to resolve intractable disagreements.

I mean "product" here pretty broadly – anything that somebody is actually going to use. It could be a literal app or widget, or an event, or a set of community norms, or a philosophical idea. You might literally sell it or just use it yourself. But I think there is something helpful about the "what if we were coworkers, how would we resolve this?" frame.

The important thing is "there is a group of people collaborating on it" and "there is a stakeholder who cares about it getting built."

If you're building a website, and one person thinks it should present all information very densely, and another person thinks it should be sleek and minimalist... *somehow* you need to actually decide what design philosophy to pursue. Options include (not necessarily limited to)

- Anarchy
- One person is in charge
- Two or more people come to consensus
- People have domain specializations in which one person is in charge (or gets veto power).

# Anarchy

To start with, what's wrong with the "everyone just builds what seems right to them and you hope it works out" option? Sometimes you're building a bazaar, not a cathedral, and this is actually fine. But it often results in different teams building different tools at cross purpose, wasting motion.

# One person in charge?

In a hierarchical company, maybe there's a boss. If the decision is about whether to paint a bikeshed red or blue, the boss can just say "red", and things move on.

This is less straightforward in the case of "minimalism" vs "high information density."

First, is the boss even doing any design work? What if the boss and the lead designer disagree about aesthetics? If the lead designer hates minimalism they're gonna have a bad time.

Maybe the boss trusts the lead designer enough to differ to them on aesthetics. Now the lead designer is the decision maker. This is an improvement, but just punts the problem down one level. If the lead designer is Just In Charge, a few things can still go wrong:

### Other workers don't actually understand minimalism

"Minimalist websites" and "information dense websites" are designed very differently. This filters into lots of small design decisions. Sometimes you can solve this with a comprehensive style guide. But those are a lot of work to create. And if you're a small startup (or a small team within a larger company), you may not have have the resources for that. It'd be nice if your employees just *actually understood* minimalism so they could build good minimalist components.

### The lead designer is wrong

Sometimes the boss's aesthetic isn't locally optimal, and this actually needs to be pointed out. If lead-designer Alice says "we're building a minimalist website" it might be important for another engineer or designer to say "Alice, you're making weird tradeoffs for minimalism that are harming the user experience."

Alice might think "Nah, you're wrong about those tradeoffs. Minimalism is great and history will bear me out on this." But Alice might also respect Bob's opinion enough to *want* to come to some kind of principled resolution. If Bob's been right about similar things before, what should Alice and Bob do, if Alice *wants* to find out she's wrong – *if and only if she's actually wrong*, and that her minimalist aesthetic is harming the user experience.

**The lead designer is right, but other major stakeholders think she's wrong**

Alternately, maybe Bob thinks Alice is making bad design calls, but Alice is actually just making the right calls. Bob has rare preferences that don't overlap much with the average user, that *shouldn't* necessitate a major design overhaul.

Initially, this will look the same to both parties as the previous option.

If Alice has listened to Bob's complaints a bunch, and Alice generally respects Bob but thinks he's wrong here, at some point she needs to say "Look Bob, we just need to actually build the damn product now, we can't rehash the minimalism argument every time we build a new widget."

I think it's useful for Bob to gain the skill of saying "Okay. fine." Let go of his frustration and embrace the design paradigm.

But that's a tough skill. And meanwhile, Bob is probably going to spend a fair amount of time and energy *being annoyed* about having to build a product they're less excited about. And sometimes, Bob's work is less efficient because he doesn't understand minimalism and keeps building site-components subtly incompatible with it.

**What if there was a process by which *either* Alice would update *or* Bob would update, that both Alice and Bob considered fair?**

You might just call that process "regular debate." But the problem is that regular debate just *often doesn't work.* Alice says "We need X, because Y". Bob says "No, we need A, because B", and they somehow both repeat those points over and over without ever changing each other's mind.

This wastes loads of time, which could have been better spent building new site features if they were able to do it faster.

Even if Alice is in charge and gets final say, it's still suboptimal for Bob to have lower morale and keep making subtly wrong widgets.

And even if Bob understands that Alice is in charge, it might still be suboptimal for Bob to feel like Alice never *really* understood exactly what Bob's concerns were.

# What if there's no boss?

Maybe your "company" is just two friends in a basement doing a project together, and there isn't really a boss. In this case, the problem is much sharper – somehow you need to actually make a call.

You might solve this by deciding to appoint a decision-maker – change the situation from a "no boss" to "boss" problem. But if you were just two friends making a game together in their spare time, for fun, this might kinda suck. (If the whole point was to

make it together as friends, a hierarchical system may be fundamentally un-fun and defeat the point)

You might be doing a more serious project, where you agree that it's important to have clear coordination protocols and hierarchy. But it nonetheless feels premature to commit to "Alice is always in charge of design decisions." Especially if Bob and Alice both have reasonable design skills. And especially if it's early on in the project and they haven't yet decided what their product's design philosophy should be.

In that case, you can start with straightforward debate, or making a pros/cons list, or exploring the space a bit and hoping you come to agreement. But if you're not coming to agreement... well, you need to do *something.*

# If "regular debate" is working for you, cool.

If "just talking about the problem" is working, obviously you don't have an issue. Sometimes the boss actually just says "we're doing it this way" and it doesn't require any extensive model sharing.

If you've never run into the problem of intractable-disagreement while collaborating on something important, this blogpost is not for you. (But, maybe keep it in the back of your mind in case you *do* run into such an issue)

But working on the LessWrong team for about 1.5 years, I've run into numerous deep disagreements, and my impression is that such disagreements are common – especially in domains where you're solving a novel problem. We've literally argued a bunch about minimalism, which isn't an especially unique design decision. We've also had much weirder disagreements about integrity and intellectual progress and AI timelines and more.

We've resolved many (although not all) of those disagreements. In many cases, doublecrux has been helpful as a framework.

# What's Doublecrux again?

If you've made it this far, presumably it seems useful to have *some* kind of process-for-consensus that works better than whatever you and your colleagues were doing by default.

**Desiderata that I personally have for such a process:**

- Both parties can agree that it's worth doing
- It should save more time than it costs (or produce value commensurate with the time you put in)
- It works even when both parties have different frames or values
- If necessary, it untangles confused questions, and replaces them with better ones
- If necessary, it untangles confused goals, and replaces them with better ones
- If people are disagreeing because of aesthetic differences like "what it beautiful/good/obviously-right", it provides a framework wherein people can actually change their mind about "what is beautiful and good and right."

- Ultimately, it lets you "get back to work", and actually build the damn product, confident that you are going about it the right way.

[Many of these goals were not assumptions I started with. They're listed here because I kept running into failures relating to each one. Over the past 2 years I've had some success with each of those points]

Importantly, it's not necessarily needed for such a process to answer the original question you asked. In the context of building a product, what's important is that you figure out a model of the world which you both agree on, which informs which actions to take.

Doublecrux is a framework that I've found helpful for the above concerns. But I think I'd consider it a win for this essay if I've at least clarified why it's desirable to have *some* such system. I share Duncan's belief that it's more promising to repair or improve doublecrux than to start from scratch. But if you'd rather start from scratch, that's cool.

**Components of Doublecrux – Cognitive Motions vs Attitudes**

There are two core concepts behind the doublecrux framework:

- *A set of cognitive motions:*
  - Looking for the cruxes of your beliefs, and asking what empirical observations would change your mind about them. (Recursing until you find a crux you and your partner both share, the "doublecrux")
- *A set of attitudes*
  - Epistemic humility
    - "*maybe I'm the wrong one*"
  - Good faith
    - "*I trust my partner to be cooperating with me*"
  - Belief that objective reality is real
    - "*there's an actual right answer here, and it's better for each of us if we've both found it*"
  - Earnest curiosity

Of those, I think the set of attitudes is more important than the cognitive motions. If the "search for cruxes and empirical tests" thing isn't working, but you have the four attitudes, you can probably find other ways to make progress. Meanwhile, if you don't each have those four attitudes, you don't have the foundations necessary to doublecrux.

**Using language for truthseeking, not politics**

But I think the cognitive motions are helpful, for this reason: much of human language is by default *politics* rather than *truthseeking*. "Regular debate" often reinforces the use of language-as-politics, which actives brain modules that are optimizing to win, which involves strategic blindness. (I mean something a bit nuanced by "politics" here, beyond scope of this post. But basically, optimizing beliefs and words for how you fit into the social landscape, rather than optimizing for what corresponds to objective reality).

The "search for empirical tests is and cruxes-of-beliefs" motion is designed to keep each participant's brain in a "language-as-truthseeking" mode. If you're asking

yourself "why would *I* change my mind?", it's more natural to be honest to yourself and your partner than if you're asking "how can I change *their* mind?"

Meanwhile, the focus on mutual, opposing cruxes keeps things *fruitful*. Disagreement is more interesting and useful than agreement – it provides an opportunity to actually learn. If people are doing language-as-politics, then disagreement is a red flag that you are on opposing sides and might be threatening each other (which might either prompt you to fight, or prompt you to "agree to disagree", preserving the social fabric by sweeping the problem under the rug).

But *if* you can both trust that everyone's truthseeking, then you can drill directly into disagreements without worrying about that, optimizing for learning, and then for building a shared model that lets you actually make progress on your product.

**Trigger Action Plans**

Knowing this is all well and good, but what might this translate into in terms of actions?

If happen to have a live disagreement *right now*, maybe you can try doublecrux. But if not, what circumstances should prompt

I've found the "Trigger Action Plan" framework useful for this sort of thing, as a basic rationality building-block skill. If you notice an unhelpful conversational pattern, you can build an association where you take some particular action that seems useful in that circumstance. (Sometimes, the generic trigger-action of "notice something unhelpful is happening ----> stop and *think*" is good enough)

In this case, a trigger-action I've found useful is:

TRIGGER: Notice that we've been arguing awhile, and someone has just repeated the same argument they said a little while ago (for the second, or especially third time)

ACTION: Say something like: "Hey, I notice that we've been repeating ourselves a bit. I feel like conversation is kinda going in circles...." followed by either "Would you be up for trying to formally doublecrux about this?" or following Duncan's vaguer suggestions about how to unilaterally improve a conversation (depending on how much shared context you and your partner have).

# Summary

- Intractable disagreements don't always matter. But if you're trying to build something together, and disagreeing substantially about how to go about it, you will need some way to resolve that disagreement.
- Hierarchy can obviate the need for resolution *if* the disagreement is simple, and if everyone agrees to respect the boss's decision.
- If the disagreement has persisted awhile and it's still wasting motion, at the very least it's probably useful to do *something* differently. In particular, if you've been repeating the
- Doublecrux is a particular framework I've found helpful for resolving intractable disagreements (when they are important enough to invest serious energy and time into). It focuses the conversation into "truthseeking" mode, and in particular strives to avoid "political mode"

# What would be the signs of AI manhattan projects starting? Should a website be made watching for these signs?

# Self-experiment Protocol: Effect of Chocolate on Sleep

# Epistemic Status

Preregistering an experiment, not very confident in the design.

Because it is important to decide how the data will be analyzed before collecting it, I'm posting my planned experiment here, may modify the design based on feedback (I'll post an update if I do), and then I will start the experiment, and finally I'll post the results when it's finished.

All feedback is appreciated!

# Introduction and Goals

I often suffer from fatigue in the early evening that leads to anxiety later at night, which makes it difficult to sleep. The purpose of this experiment is to determine whether eating a small amount of chocolate to counteract fatigue in the early evening affects my sleep, and if so, how. I could easily imagine the effect on sleep being either beneficial (counteract the pattern of fatigue and anxiety that makes it hard to sleep) or harmful (the theobromine might keep me awake).

Because I'd rather not add the fat and sugar that comes with chocolate to my diet, the effect on sleep needs to be appreciable for it to be worthwhile. I have fairly arbitrarily decided that the chocolate is worth it iff it causes me to fall asleep at least 15 minutes earlier than I otherwise would.

# Experimental Protocol

Every work day evening when I go directly home from work and do not have a social event planned, I will get home, make any necessary adjustments to the air conditioning (temperature affects my sleep a lot, and I want to make sure I don't accidentally bias the results by setting the AC differently depending on whether I have chocolate), and then with probability 1/2 (decided by a coin flip or die roll) eat a square of dark chocolate.

I will use a spreadsheet to track which nights I did and did not eat a square of chocolate, and I will use sleep data from a Fitbit Blaze smartwatch to record what time I fall asleep each night. The Blaze uses heart rate and movement data to decide what time I fall asleep (I don't need to actively tell it when I go to bed). In my experience, the time it says I fell asleep generally matches my subjective memory of when I fell asleep (not when I went to bed).

# Statistical Analysis

I care about the magnitude of the effect, not just the direction, and my goal is to make the correct decision rather than to get published, so testing whether the observed effect is statistically significantly different from 0 is not very useful. While it would be a mistake to keep going until I get the result I want, if the results are close (in either direction), it will be worth gathering more data.

Because I have arbitrarily picked an absolute effect size of -15 minutes (I fall asleep 15 minutes sooner than I otherwise would), I will be focusing on determining which side of that cut-off the effect is. To do so, I plan on analyzing the data as follows:

1. Run the experiment until I have used up one bag of chocolate (15 individually-wrapped squares).
2. Calculate the standard error of the difference of sample means (I am not assuming equal variance in the two samples).
3. If the observed effect is at least one standard error away from the -15 minute cut-off (my sleep is inconsistent enough that I expect this condition probably won't be met at this point), stop the experiment.
4. Otherwise, keep going until either the stop condition from step 3 is met or the standard error is less than 5 minutes. If I end with a standard error less than 5 minutes and the observed effect is within 5 minutes of -15 minutes, I will consider the experiment inconclusive.

# The Right Way of Formulating a Problem?

David Chapman Writes,

> Finding a good formulation for a problem is often most of the work of solving it.

I agree with this intuitively, and I feel like I have seen this principle at work in my own work and in the problems I have tried to solve. However, when I try to convince others of this idea, I struggle to find examples that they can connect with or that they find compelling.

I suspect that programmers find this idea appealing because we routinely work with formal systems, and all of us know the experience of making a minor change in perspective and seeing an impossible problem turn into an easy one. So I'm most interested in examples that have nothing to do with code, examples that a lay audience would be able to grasp.

I would be *particularly* interested in examples from the history of science or medicine, if anyone can think of some. Scott and Scurvy is the only example I currently know of, and while interesting, does not seem like a perfect fit.

Much appreciated!

# Self-consciousness wants to make everything about itself

This is a linkpost for https://unstableontology.com/2019/07/03/self-consciousness-wants-to-make-everything-about-itself/

Here's a pattern that shows up again and again in discourse:

A: This thing that's happening is bad.

B: Are you saying I'm a bad person for participating in this? How mean of you! I'm not a bad person, I've done X, Y, and Z!

It isn't always this explicit; I'll discuss more concrete instances in order to clarify. The important thing to realize is that A is pointing at a concrete problem (and likely one that is concretely affecting them), and B is changing the subject to be about B's own self-consciousness. Self-consciousness wants to make everything about itself; when some topic is being discussed that has implications related to people's self-images, the conversation frequently gets redirected to be *about* these self-images, rather than the concrete issue. Thus, problems don't get discussed or solved; everything is redirected to being about maintaining people's self-images.

## Tone arguments

A tone argument criticizes an argument not for being incorrect, but for having the wrong tone. Common phrases used in tone arguments are: "More people would listen to you if...", "you should try being more polite", etc.

It's clear why tone arguments are epistemically invalid. If someone says X, then X's truth value is independent of their tone, so talking about their tone is changing the subject. (Now, if someone is saying X in a way that breaks epistemic discourse norms, then defending such norms is epistemically sensible; however, tone arguments aren't about epistemic norms, they're about people's feelings).

Tone arguments are about people protecting their self-images when they or a group they are part of (or a person/group they sympathize with) is criticized. When a tone argument is made, the conversation is no longer about the original topic, it's about how talking about the topic in certain ways makes people feel ashamed/guilty. Tone arguments are a key way self-consciousness makes everything about itself.

Tone arguments are practically always in bad faith. They aren't made by people trying to help an idea be transmitted to and internalized by more others. They're made by people who want their self-images to be protected. Protecting one's self-image from the truth, by re-directing attention away from the epistemic object level, is acting in bad faith.

## Self-consciousness in social justice

A documented phenomenon in social justice is "white women's tears". Here's a case study (emphasis mine):

> A group of student affairs professionals were in a meeting to discuss retention and wellness issues pertaining to a specific racial community on our campus. As the dialogue progressed, **Anita, a woman of color, raised a concern about the lack of support and commitment to this community from Office X** (including lack of measurable diversity training, representation of the community in question within the staff of Office X, etc.), **which caused Susan from Office X, a White woman, to feel uncomfortable. Although Anita reassured Susan that her comments were not directed at her personally, Susan began to cry while responding that she "felt attacked".** Susan further added that: she donated her time and efforts to this community, and even served on a local non-profit organization board that worked with this community; she understood discrimination because her family had people of different backgrounds and her closest friends were members of this community; she was committed to diversity as she did diversity training within her office; and the office did not have enough funding for this community's needs at that time.
>
> Upon seeing this reaction, Anita was confused because although her tone of voice had been firm, she was not angry. From Anita's perspective, the group had come together to address how the student community's needs could be met, which partially meant pointing out current gaps where increased services were necessary. Anita was very clear that she was critiquing Susan's office and not Susan, as Susan could not possibly be solely responsible for the decisions of her office.
>
> **The conversation of the group shifted at the point when Susan started to cry. From that moment, the group did not discuss the actual issue of the student community. Rather, they spent the duration of the meeting consoling Susan, reassuring her that she was not at fault.** Susan calmed down, and publicly thanked Anita for her willingness to be direct, and complimented her passion. **Later that day, Anita was reprimanded for her 'angry tone,' as she discovered that Susan complained about her "behavior" to both her own supervisor as well as Anita's supervisor.** Anita was left confused by the mixed messages she received with Susan's compliment, and Susan's subsequent complaint regarding her.

The key relevance of this case study is that, while the conversation was originally about the issue of student community needs, it became about Susan's self-image. Susan made everything about her own self-image, ensuring that the actual concrete issue (that her office was not supporting the racial community) was not discussed or solved.

# Shooting the messenger

In addition to crying, Susan also shot the messenger, by complaining about Anita to both her and Anita's supervisors. This makes sense as ego-protective behavior: if she wants to maintain a certain self-image, she wants to discourage being presented with information that challenges it, and also wants to "one-up" the person who challenged her self-image, by harming that person's image (so Anita does not end up looking better than Susan does).

Shooting the messenger is an ancient tactic, deployed especially by powerful people to silence providers of information that challenges their self-image. Shooting the messenger is asking to be lied to, using force. Obviously, if the powerful person actually wants information, this tactic is counterproductive, hence the standard advice to not shoot the messenger.

# Self-consciousness as privilege defense

It's notable that, in the cases discussed so far, self-consciousness is more often a behavior of the privileged and powerful, rather than the disprivileged and powerless. This, of course, isn't a hard-and-fast rule, but there certainly seems to be a relation. Why is that?

Part of this is that the less-privileged often *can't get away with* redirecting conversations by making everything about their self-image. People's sympathies are more often with the privileged.

Another aspect is that privilege is largely *about* being rewarded for one's identity, rather than one's works. If you have no privilege, you have to actually do something concretely effective to be rewarded, like cleaning. Whereas, privileged people, almost by definition, get rewarded "for no reason" other than their identity.

Maintenance of a self-image makes less sense as an individual behavior than as a collective behavior. The phenomenon of bullshit jobs implies that much of the "economy" is *performative*, rather than about value-creation. While almost everyone can pretend to work, some people are better at it than others. The best people at such pretending are those who look the part, and who maintain the act. That is: privileged people who maintain their self-images, and who tie their self-images to their collective, as Susan did. (And, to the extent that e.g. school "prepares people for real workplaces", it trains such behavior.)

Redirection away from the object level isn't merely about defending self-image; it has the effect of causing issues not to be discussed, and problems not to be solved. Such effects maintain the local power system. And so, power systems encourage people to tie their self-images with the power system, resulting in self-consciousness acting as a defense of the power system.

Note that, while less-privileged people do often respond negatively to criticism from more-privileged people, such responses are more likely to be based in fear/anger rather than guilt/shame.

# Stop trying to be a good person

At the root of this issue is the desire to maintain a narrative of being a "good person". Susan responded to the criticism of her office by listing out reasons why she was a "good person" who was against racial discrimination.

While Anita wasn't actually accusing Susan of racist behavior, it is, empirically, likely that some of Susan's behavior *is* racist, as implicit racism is pervasive (and, indeed, Susan silenced a woman of color speaking on race). Susan's implicit belief is that there is such a thing as "not being racist", and that one gets there by passing some threshold of being nice to marginalized racial groups. But, since racism is a *structural*

issue, it's quite hard to actually stop participating in racism, without going and living in the woods somewhere. In societies with structural racism, ethical behavior requires skillfully and consciously reducing harm *given* the fact that one is a participant in racism, rather than washing one's hands of the problem.

What if it isn't actually possible to be "not racist" or otherwise "a good person", at least on short timescales? What if almost every person's behavior is morally depraved a lot of the time (according to their standards of what behavior makes someone a "good person")? What if there are bad things that *are* your fault? What would be the right thing to do, then?

Calvinism has a theological doctrine of total depravity, according to which every person is utterly unable to stop committing evil, to obey God, or to accept salvation when it is offered. While I am not a Calvinist, I appreciate this teaching, because quite a lot of human behavior is simultaneously unethical and hard to stop, and because accepting this can get people to stop chasing the ideal of being a "good person".

If you accept that you are irredeemably evil (with respect to your *current* idea of a good person), then there is no use in feeling self-conscious or in blocking information coming to you that implies your behavior is harmful. The only thing left to do is to *steer in the right direction*: make things around you better instead of worse, based on your intrinsically motivating *discernment* of what is better/worse. Don't try to be a good person, just try to make nicer things happen. And get more foresight, perspective, and cooperation as you go, so you can participate in steering bigger things on longer timescales using more information.

Paradoxically, in accepting that one is irredeemably evil, one can start accepting information and steering in the right direction, thus developing *merit*, and becoming a better person, though still not "good" in the original sense. (This, I know from personal experience)

(See also: What's your type: Identity and its Discontents; Blame games; Bad intent is a disposition, not a feeling)

# Largest open collection quotes about AI

This is a linkpost for
https://drive.google.com/file/d/1bnuiXukz6tb7CnB6_bXBj8mh7uaL2ZcR/view

I apologize for my bad English, this is not my native language. And probably I will make some mistakes when posting.

For over 2 years I have been reading materials on the topic of AI Safety. I don't have the appropriate education, cognitive abilities, knowledge. I do not even have time to learn the language. So I didn't hope to do something useful myself.

But once I tried to systematize quotations from one show in order to understand when the experts represented there are waiting for AGI and how likely they consider the extinction of humanity.

I thought it would be interesting to do so with the rest of the experts.
In addition, I have already seen and studied with interest such collections of quotes. It seemed to me that the best thing I could do was try to do something similar.

Therefore, I began to collect quotes from people who can be attributed to the experts. It turned out to be harder than I thought.

I have compiled a table with quotes from more than 800 experts. I tried not to distort the opinion of forecasters and simply copied from sources, sometimes deleting or slightly editing. My edits can be recognized by square brackets :)

1) The first column of the table is the name of the expert.
2) The second column is the year of the forecast. The table is built in chronological order.
3) The third column is the predicted time for AGI. Unfortunately, most people did not speak directly about time and probability. Because of this, many quotes came out rather vague. For example, “Machines are very far from being intelligent” or “And we can reach it in a close time”.
4) The fourth column is an opinion about Takeoff Speed. About how much progress will be accelerated after create of AGI.
5) The fifth column is the expert's opinion about the future of mankind with AGI. Choosing a quote here was the hardest. Most of all I was interested in the risk of extinction or serious shocks due to AI, and I tried to provide quotes that most fully reveal this particular topic.
6) The sixth column indicates the source of the quote.

That is, to the right of the forecaster's name, you can find out the date of the given quotes, his opinion about the time of the creation of AI, about the intellectual explosion and about the future of humanity, as well as get acquainted with the source.

Of course, cases where the expert spoke on the topic of time, the speed of self-improvement and the influence of AI in the framework of one material are quite rare. Therefore many cells are left empty.

I had to give several quotes per person, sometimes they were separated for years and even decades.

Since all the quotes are given in chronological order, the opinions of some people are “scattered” in the table.

For example, Gwern spoke about the future of mankind in 2010, about the growth of AI in 2014 and about the forecasts for the emergence of AI in 2018.
However, you can simply use search.

In addition, sometimes one person has already made a certain forecast but later changed or expanded his opinion. I tried to take into account such quotes.

I also reviewed anonymous expert interviews and indicated them. If the general list of respondents was known, I cited them as well.

It was difficult to decide who should be considered an expert and what quotes should be included in the work.

I had to make controversial decisions. The table includes a lot of people who are entrepreneurs but may have insights on advanced research. There are several futurists and philosophers in the table. There are writers like Clark and Vinge, whose opinion seems important to me.

I have a version of this work without chronological separation, where the quotes are more grouped by name. Perhaps someone will be more convenient.

It is difficult to draw conclusions from the work. The absolute majority of experts did not talk about exact dates and did not indicate the probability of their predictions.

I can only say that most forecasters do not expect AI in the near future, do not expect IE and seem optimistic.

In addition, it seemed to me that in the twentieth century the leading experts were on average more pessimistic: Turing, Wiener, I. J. Good, Fredkin, Shannon, Moravec, etc.

Young researchers are on average more optimistic than older ones - even in the field of AI Safety, where on average there are naturally more concerned people.

I think that to confirm almost any views you can find the opinion of a respected expert.

I really hope that for someone my work will be useful and interesting. Criticism and additions are welcome.

# Black hole narratives

Related to: Self-consciousness wants to make everything about itself (by jessicata), UNIVERSAL LOVE, SAID THE CACTUS PERSON (by SSC) and Ms. Blue, meet Mr. Green

Edit (Jul 17, 2019): The concept I'm describing is basically cognitive fusion.

There's an idea that I think is really important to understand and to try to communicate. I think Jessica in her post touched on it a little, but that touch inspired me to give it a shot myself. However, instead of focusing on social dynamics, I'm going to focus on an individual's internal world and interpretation.

Here are a few examples of mental narratives that I'm going to dissect. (I borrowed a few from Jessica's post and a few from the SSC post.) See if you can find the ones that happen to you. Or see if you can understand the pattern to find ones that are not on the list.

1. This thing that's happening is bad. I wish it wouldn't happen.

2. Am I a bad person for participating in this? I think something's wrong with me.

3. I'm not a bad person! I'm definitely in the right here.

4. I'm feeling attacked. I feel like people are against me.

5. I don't feel safe here. I think they'll hurt me.

6. I feel like I'm the odd one out. I don't think I belong here.

7. I don't understand this at all. What am I missing? There's a clue here somewhere. If I can find it, then I'll understand it.

# Lenses

To try to understand the pattern and what to do with it, I'll provide four different lenses.

## Physical sensations

What does it feel like to experience these narratives? Look and see.

For me, it feels constraining. It feels like sensing a wall and then pushing away from it. Or pushing into it. It feels like a river that has hit the bank and has to turn. My body feels tighter. My awareness feels more narrow or disappears entirely. The thought grows bigger in my mind until it dominates. Until, just for a moment, it feels like my entire mental reality is described by the circumstance. By how I feel attacked. Or by how I feel like I don't belong. It stops being a thought and becomes reality.

So how does it feel when the narrative ends? To me it feels like someone was hugging me super tight and then let go. It feels like a release. An expansion. An exhale. It feels

like freedom. Freedom to reconsider. To review. To remember other sides of the problem. Freedom to forget. Freedom to move on.

# TAP

We can also think of those narratives as serving a particular purpose. Like other TAPs, they were installed there for some reason, probably a long time. Probably because someone told us. Or probably because a "bad" thing happened to us and we really wanted for it to not happen again.

Each narrative has many triggers. But once triggered, the action is pretty straight forward. And it's usually always the same: recount the narrative and listen to it. Just like when Revolio Clockberg Jr. hears about Gear Wars, he *has* to tell you all about it. And once he starts, there's no stopping it. Like clockwork.

# Movie

One of the best ways to talk about narratives is to talk about stories and movies. Movies are compelling. They draw us in. When we're in a middle of a good film, we forget who we are. We feel wholeheartedly the emotion, the action, the tension. Where is all of this going? Well... we know. It's going where the movie going. There's only one track.

And for a great movie, that's a fun track to be on. But for the narratives I've described, that's not a fun track at all. You've seen some of them thousands of times already. Do you want to see it again? All it costs is a few seconds of your time and a few drops of your sanity.

"I'm feeling attacked. Like people are against me." is a movie. It's a lousy, poorly written, utterly predictable movie. And you know exactly where it's heading. "Me against the world." Or "Look at me, I'm so independent. I don't need anyone." Or "Hah! They're all against me, but look who is winning now!" And, honestly, even that is giving that movie too much credit. Because 99.9% of the time the only place that movie is taking you to is: "I'm feeling attacked. Like people are against me. This sucks. This sucks. This sucks."

# Car

When you're gripped by one of such narratives, you're in the car. When you aren't, you're free. Get out of the car and stay out!

Q: I'm experiencing all these mental narratives. What do I do? What's the skill I need to learn to avoid them?

A: Note how that question itself is an example of the narrative / car you're trying to avoid. Drop it.

Q: I was trying hard to avoid all the mental narratives you described. But then I ran into one head first. Where and how should I put up more walls so that this wouldn't happen?

A: Stop putting up walls. Stop seeing walls. Stop trying to avoid them by creating more walls. When you're not in the car, there are no walls and there are no roads.

"Then I can’t get out of the car. I want to get out of the car. But I need help. And the first step to getting help is for you to factor my number." is being in the car. It's watching a movie about how you're stuck and how you're missing something. It's running a TAP that says: IF I don't understand something THEN I'm unsafe until I do. It's screaming at the cactus person, demanding an answer.

As long as you think there's a car or that you have something to do about it, you can't get out of it. Stop trying to get out of the car and get out of the car.

# The invitation

The invitation is to notice when these narratives arise. Just notice. You don't have to do anything about them. You don't have to stop them or avoid them. Just notice when one starts. Notice when you're in it. Notice when it ends. Notice when it's not there.

And if on one of those occasions you decide to pull over that car and get out... well, I promise you the car is not going to be upset.

# Prereq: Cognitive Fusion

In a post by Kaj Sotala, he introduces the very useful idea of cognitive fusion.

> Cognitive fusion is a term from Acceptance and Commitment Therapy (ACT), which refers to a person “fusing together” with the content of a thought or emotion, so that the content is experienced as an objective fact about the world rather than as a mental construct. The most obvious example of this might be if you get really upset with someone else and become convinced that something was *all their fault* (even if you had actually done something blameworthy too). In this example, your anger isn’t letting you see clearly, and you can’t step back from your anger to question it, because you have become “fused together” with it and experience everything in terms of the anger’s internal logic.

You can become fused to an emotion, a voice in your head, a political view, and experience it to "just be true". I see this as a similar sort of fusion I hear musician talk about, where after years of practice their instrument begin to feel like a part of their body. They aren't "using their index finger to press the black note on a piano" they are "just playing G". This is analogous to being so caught up in your own anger that your partner is "just wrong and terrible" as opposed to "it sorta looks like you intentionally did something to annoy me and I'm worried about if you'll do this again in the future." (or whatever the actual case is)

Sometimes I think of there being a general fusion process where the brain collapses levels of inference. All of the steps that go into a given physical motion or thought process get compressed into a single dot. The thought process will be experienced as "just true" and the physical motion will be experience as an atomic action available to you. Sometimes you can "uncompress" the chain, and sometimes you can't.

Problems can arise when you fuse to a thought or emotion that doesn't have an accurate view of the world, and you unknowingly take it's broken map as the territory.

## Isn't this just "Don't make assumptions"?

Not quite, though it is similar. Assumptions don't really capture the more general fusing process that you can also see with physical movement. "I can't believe that you just assumed you start off on your left foot when making a layup!" Nah, doesn't feel right. But the main reason I prefer to talk in terms of fusion is that "fusion" makes me focus on the process of attaching to something, while "assumptions" makes me focus on the object being attached to.

It's easier to see this difference when the thought being fused to (or the claim being assumed) is "obviously" wrong, or at least obvious to one who isn't fused to it. The assumption frame makes me feel like my work is done when I find the other persons "dumb" assumption. Point it out with a pithy "Checkmate [outgroup]" and move on. The fusion frame leads me to ask "How did they get fused to this in the first place? How might I help them defuse from it?" By focusing on the process of attachment (fusion) I can appreciate how common it is to fuse to something and how hard it can be to defuse. When I focus on the object of attachment (assumption) I'm mostly thinking about just how *stupid* it is and how I can't believe that anyone would be *dumb* enough to fall for this, and I most certainly don't believe anything that stupid....

And so it goes. Thinking of some behavior as a "dumb mistake" makes you *more likely to not notice when you engage in it*. Some thoughts take moments to defuse from. Others take a lifetime. Sometimes you fuse to things, and you'd be wise to learn how it works rather than to ridicule it.

As you may have guessed, later parts of this sequence will talk about what can happen when you fuse with language. For now, just remember what fusion is, and treat it with respect.

> "Modern man can't see God because he doesn't look low enough."
>
> -- Carl Jung

# On the purposes of decision theory research

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

Following the examples of Rob Bensinger and Rohin Shah, this post will try to clarify the aims of part of my research interests, and disclaim some possible misunderstandings about it. (I'm obviously only speaking for myself and not for anyone else doing decision theory research.)

I think decision theory research is useful for:

1. Gaining information about the nature of rationality (e.g., is "realism about rationality" true?) and the nature of philosophy (e.g., is it possible to make real progress in decision theory, and if so what cognitive processes are we using to do that?), and helping to solve the problems of normativity, meta-ethics, and metaphilosophy.
2. Better understanding potential AI safety failure modes that are due to flawed decision procedures implemented in or by AI.
3. Making progress on various seemingly important intellectual puzzles that seem directly related to decision theory, such as free will, anthropic reasoning, logical uncertainty, Rob's examples of counterfactuals, updatelessness, and coordination, and more.
4. Firming up the foundations of human rationality.

To me, decision theory research is *not* meant to:

5. Provide a correct or normative decision theory that will be used as a specification or approximation target for programming or training a potentially superintelligent AI.
6. Help create "safety arguments" that aim to show that a proposed or already existing AI is free from decision theoretic flaws.

To help explain 5 and 6, here's what I wrote in a previous comment (slightly edited):

> One meta level above what even UDT tries to be is decision theory (as a philosophical subject) and one level above that is metaphilosophy, and my current thinking is that it seems bad (potentially dangerous or regretful) to put any significant (i.e., superhuman) amount of computation into anything *except* doing philosophy.
>
> To put it another way, any decision theory that we come up with might have some kind of flaw that other agents can exploit, or just a flaw in general, such as in how well it cooperates or negotiates with or exploits other agents (which might include how quickly/cleverly it can make the necessary commitments). Wouldn't it be better to put computation into trying to find and fix such flaws (in other words, coming up with better decision theories) than into any particular object-level decision theory, at least until the superhuman philosophical computation itself decides to start doing the latter?

Comparing my current post to Rob's post on the same general topic, my mentions of 1, 2, and 4 above seem to be new, and he didn't seem to share (or didn't choose to emphasize) my concern that decision theory research (as done by humans in the foreseeable future) can't solve decision theory in a definitive enough way that would obviate the need to make sure that any potentially superintelligent AI can find and fix decision theoretic flaws in itself.

# Job description for an independent AI alignment researcher

This is the job description that I've written for myself in order to clarify what I'm supposed to be doing.

I'm posting it here in order to get feedback on my understanding of the job. Also, if you're thinking of becoming an independent researcher, you might find it useful to know what it takes.

---

## Admin

Job Title: Independent AI alignment researcher

Location: anywhere (in my case: Kagoshima, Japan)

Reports To: nobody (in a sense: funders, mentors, givers of feedback)

Position Status: not applicable

## Responsibilities

- Define AI alignment research projects. Includes finding questions, gauging their significance and devising ways to answer them.
- Execute research projects by reading, thinking and experimenting.
- Write and publish results in the form of blog entries, contributions to discussions and conferences (conference paper, presentation, poster), journal articles, public datasets, software.
- Solicit feedback and use it to improve processes and results.
- Find potential junior (in the sense of (slightly) less experienced in the field) researchers and help them grow.
- Help other researchers with their work.
- Make sure that the money doesn't run out.
- Any other activities required by funders or givers of feedback.

## Hiring requirements

Entry level:

- Strong desire to do good for the world by contributing to AI alignment.
- Undergrad degree or equivalent skill level in computer science, maths or machine learning. Includes having researched, written and presented a scientific paper or thesis.
- Ability to define, plan and complete novel projects with little supervision.
- Ability to collaborate remotely.
- Initiative.
- Discipline.

- Ability to speak and write clearly.
- Ability to identify and close gaps in knowledge or skills.
- Ability to write job or funding applications.
- Ability to figure out things that are usually taken care of for employees: taxes, insurance, payments, bookkeeping, budgeting.
- Ability to deal with uncertainty and financial stress.

---

# Resources used

- Manager Tools: Writing a Job Description
- Booth et al.: The Craft of Research

# Diversify Your Friendship Portfolio

In investing, a common piece of advice recommends a diversified portfolio - in other words, having investments from a wide range of areas to avoid a crash in one area wiping out all your investments. Similarly, I have found that it can be beneficial to have a diversified *friendship* portfolio - in other words, friends across a wide range of communities or social circles.

For instance, I have a group of friends that I've known mostly since elementary/middle school, a group of friends that I know from the rationalist and EA community in Berkeley, a group of friends that I know from playing certain games competitively... there is some overlap between these groups, but by and large they are separate, and this can be quite valuable.

Because I'm active in multiple circles, if strange things are happening in one community I can step into another. If I need advice on a sensitive situation, I have people who know me well and aren't close to the matter to draw upon. Further, having these sorts of resources and perspectives available can open up options that might otherwise be difficult - if, for instance, I decided I could no longer be a part of one of those communities, leaving wouldn't be the end of my social life.

I sometimes see people -- in the rationalist community or elsewhere -- putting "all their eggs in one basket" when it comes to friendship, and I think that can often lead to pitfalls. If all your friends are from work, what happens if you leave your job? If all your friends are from a certain hobby, what happens if you get bored of it? If all your friends are from a certain social scene, what happens if there's a bunch of drama and that community splits? Having other social connections and communities to interact with can really help with such scenarios.

Lastly, I want to point out that having a range of friend groups can be a useful insulator against bad ideas. [1] Sometimes strange and unwelcome fads can spread across a community, and if that's the case it can suddenly become a less appealing place. If you have several friend groups to choose from, such things are much easier to "ride out" than if your whole social circle is suddenly into whatever new weird thing. Further, being able to run new stuff by people you know in other communities can serve as a good check on groupthink - being able to say "hey, a bunch of my friends in Berkeley are getting into X, does that make much sense to you?" can be quite a useful reality check!

[1] One caveat to this -- if you notice that certain friends or communities constantly seem to drag you in bad directions, it might well be time to move away from them!

# Should I wear wrist-weights while playing Beat Saber?

I recently started playing the VR game Beat Saber as a form of exercise. It involves waving your arms around a bunch.

I *also* got 1.5 pound wrist weights to use while playing, to increase the degree of workout.

Since then, someone made a vague claim about this potentially being damaging for joints, or something. I'm curious if anyone has a clear sense of whether and why wrist-weights would be beneficial or harmful.

# Dialogue on Appeals to Consequences

This is a linkpost for https://unstableontology.com/2019/07/18/dialogue-on-appeals-to-consequences/

[note: the following is essentially an expanded version of this LessWrong comment on whether appeals to consequences are normative in discourse. I am exasperated that this is even up for debate, but I figure that making the argumentation here explicit is helpful]

Carter and Quinn are discussing charitable matters in the town square, with a few onlookers.

Carter: "So, this local charity, People Against Drowning Puppies (PADP), is nominally opposed to drowning puppies."

Quinn: "Of course."

Carter: "And they said they'd saved 2170 puppies last year, whereas their total spending was $1.2 million, so they estimate they save one puppy per $553."

Quinn: "Sounds about right."

Carter: "So, I actually checked with some of their former employees, and if what they say and my corresponding calculations are right, they actually only saved 138 puppies."

Quinn: "Hold it right there. Regardless of whether that's true, it's bad to say that."

Carter: "That's an appeal to consequences, well-known to be a logical fallacy."

Quinn: "Is that really a fallacy, though? If saying something has bad consequences, isn't it normative not to say it?"

Carter: "Well, for my own personal decisionmaking, I'm broadly a consequentialist, so, yes."

Quinn: "Well, it follows that appeals to consequences are valid."

Carter: "It isn't *logically* valid. If saying something has bad consequences, that doesn't make it false."

Quinn: "But it is decision-theoretically *compelling*, right?"

Carter: "In theory, if it could be proven, yes. But, you haven't offered any proof, just a statement that it's bad."

Quinn: "Okay, let's discuss that. My argument is: PADP is a good charity. Therefore, they should be getting more donations. Saying that they didn't save as many puppies as they claimed they did, in public (as you just did), is going to result in them getting fewer donations. Therefore, your saying that they didn't save as many puppies as they claimed to is bad, and is causing more puppies to drown."

Carter: "While I could spend more effort to refute that argument, I'll initially note that you only took into account a single effect (people donating less to PADP) and neglected other effects (such as people having more accurate beliefs about how charities work)."

Quinn: "Still, you have to admit that my case is plausible, and that some onlookers are convinced."

Carter: "Yes, it's plausible, in that I don't have a full refutation, and my models have a lot of uncertainty. This gets into some complicated decision theory and sociological modeling. I'm afraid we've gotten sidetracked from the relatively clear conversation, about how many puppies PADP saved, to a relatively unclear one, about the decision theory of making actual charity effectiveness clear to the public."

Quinn: "Well, sure, we're into the weeds now, but this is important! If it's actually bad to say what you said, it's important that this is widely recognized, so that we can have fewer... *mistakes* like that."

Carter: "That's correct, but I feel like I might be getting trolled. Anyway, I think you're shooting the messenger: when I started criticizing PADP, you turned around and made the criticism about me saying that, directing attention against PADP's possible fraudulent activity."

Quinn: "You still haven't refuted my argument. If you don't do so, I win by default."

Carter: "I'd really rather that we just outlaw appeals to consequences, but, fine, as long as we're here, I'm going to do this, and it'll be a learning experience for everyone involved. First, you said that PADP is a good charity. Why do you think this?"

Quinn: "Well, I know the people there and they seem nice and hardworking."

Carter: "But, they said they saved over 2000 puppies last year, when they actually only saved 138, indicating some important dishonesty and ineffectiveness going on."

Quinn: "*Allegedly*, according to *your calculations*. Anyway, saying that is bad, as I've already argued."

Carter: "Hold up! We're in the middle of *evaluating* your argument that saying that is bad! You can't use the conclusion of this argument in the course of proving it! That's circular reasoning!"

Quinn: "Fine. Let's try something else. You said they're being *dishonest*. But, I know them, and they wouldn't tell a lie, consciously, although it's possible that they might have some motivated reasoning, which is totally different. It's really *uncivil* to call them dishonest like that. If everyone did that with the willingness you had to do so, that would lead to an all-out rhetorical *war*..."

Carter: "God damn it. You're making another appeal to consequences."

Quinn: "Yes, because I think appeals to consequences are normative."

Carter: "Look, at the start of this conversation, your argument was that saying PADP only saved 138 puppies is bad."

Quinn: "Yes."

Carter: "And now you're in the course of arguing that it's bad."

Quinn: "Yes."

Carter: "Whether it's bad is a matter of fact."

Quinn: "Yes."

Carter: "So we have to be trying to get the right answer, when we're determining whether it's bad."

Quinn: "Yes."

Carter: "And, while appeals to consequences may be decision theoretically compelling, they don't directly bear on the facts."

Quinn: "Yes."

Carter: "So we shouldn't have appeals to consequences in conversations about whether the consequences of saying something is bad."

Quinn: "Why not?"

Carter: "Because we're trying to get to the truth."

Quinn: "But aren't we also trying to avoid all-out rhetorical wars, and puppies drowning?"

Carter: "If we want to do those things, we have to do them *by* getting to the truth."

Quinn: "The truth, *according to your opinion*-"

Carter: "God damn it, you just keep trolling me, so we never get to discuss the actual facts. God damn it. Fuck you."

Quinn: "Now you're just spouting insults. That's really irresponsible, given that I just accused you of doing something bad, and causing more puppies to drown."

Carter: "You just keep controlling the conversation by OODA looping faster than me, though. I can't refute your argument, because you appeal to consequences again in the middle of the refutation. And then we go another step down the ladder, and never get to the truth."

Quinn: "So what do you expect me to do? Let you insult well-reputed animal welfare workers by calling them *dishonest*?"

Carter: "*Yes*! I'm modeling the PADP situation using *decision-theoretic models*, which require me to *represent* the knowledge states and optimization pressures exerted by different agents (both conscious and unconscious), including when these optimization pressures are towards deception, and even when this deception is unconscious!"

Quinn: "Sounds like a bunch of nerd talk. Can you speak more plainly?"

Carter: "I'm modeling the actual facts of how PADP operates and how effective they are, not just how well-liked the people are."

Quinn: "Wow, that's a strawman."

Carter: "Look, how do you think arguments are supposed to work, exactly? Whoever is best at claiming that their opponent's argumentation is evil wins?"

Quinn: "Sure, isn't that the same thing as who's making better arguments?"

Carter: "If we argue by proving our statements are true, we reach the truth, and thereby reach the good. If we argue by proving each other are being evil, we don't reach the truth, nor the good."

Quinn: "In this case, though, we're talking about *drowning puppies*. Surely, the good in this case is causing fewer puppies to drown, and directing more resources to the people saving them."

Carter: "That's under contention, though! If PADP is lying about how many puppies they're saving, they're making the epistemology of the puppy-saving field worse, leading to fewer puppies being saved. And, they're taking money away from the next-best-looking charity, which is probably more effective if, unlike PADP, they're not lying."

Quinn: "How do you know that, though? How do you know the money wouldn't go to things other than saving drowning puppies if it weren't for PADP?"

Carter: "I don't know that. My guess is that the money might go to other animal welfare charities that claim high cost-effectiveness."

Quinn: "PADP is quite effective, though. Even if your calculations are right, they save about one puppy per $10,000. That's pretty good."

Carter: "That's not even that impressive, but even if their direct work is relatively effective, they're destroying the epistemology of the puppy-saving field by lying. So effectiveness basically caps out there instead of getting better due to better epistemology."

Quinn: "What an exaggeration. There are lots of other charities that have misleading marketing (which is totally not the same thing as lying). PADP isn't singlehandedly destroying anything, except instances of puppies drowning."

Carter: "I'm beginning to think that the difference between us is that I'm anti-lying, whereas you're pro-lying."

Quinn: "Look, I'm only in favor of lying when it has good consequences. That makes me different from pro-lying scoundrels."

Carter: "But you have really sloppy reasoning about whether lying, in fact, has good consequences. Your arguments for doing so, when you lie, are made of Swiss cheese."

Quinn: "Well, I can't deductively *prove* anything about the real world, so I'm using the most relevant considerations I can."

Carter: "But you're using reasoning processes that systematically protect certain cached facts from updates, and use these cached facts to justify not updating. This was very clear when you used outright circular reasoning, to use the cached fact that denigrating PADP is bad, to justify terminating my argument that it wasn't bad to

denigrate them. Also, you said the PADP people were nice and hardworking as a reason I shouldn't accuse them of dishonesty... but, the fact that PADP saved far fewer puppies than they claimed actually casts doubt on those facts, and the relevance of them to PADP's effectiveness. You didn't update when I first told you that fact, you instead started committing rhetorical violence against me."

Quinn: "Hmm. Let me see if I'm getting this right. So, you think I have false cached facts in my mind, such as PADP being a good charity."

Carter: "Correct."

Quinn: "And you think those cached facts tend to protect themselves from being updated."

Carter: "Correct."

Quinn: "And you think they protect themselves from updates by generating bad consequences of making the update, such as fewer people donating to PADP."

Carter: "Correct."

Quinn: "So you want to outlaw appeals to consequences, so facts have to get acknowledged, and these self-reinforcing loops go away."

Carter: "Correct."

Quinn: "That makes sense from your perspective. But, why should I think my beliefs are wrong, and that I have lots of bad self-protecting cached facts?"

Carter: "If everyone were as willing as you to lie, the history books would be full of convenient stories, the newspapers would be parts of the matrix, the schools would be teaching propaganda, and so on. You'd have no reason to trust your own arguments that speaking the truth is bad."

Quinn: "Well, I guess that makes sense. Even though I lie in the name of good values, not everyone agrees on values or beliefs, so they'll lie to promote their own values according to their own beliefs."

Carter: "Exactly. So you should expect that, as a reflection to your lying to the world, the world lies back to you. So your head is full of lies, like the 'PADP is effective and run by good people' one."

Quinn: "Even if that's true, what could I possibly do about it?"

Carter: "You could start by not making appeals to consequences. When someone is arguing that a belief of yours is wrong, listen to the argument at the object level, instead of jumping to the question of whether saying the relevant arguments out loud is a good idea, which is a much harder question."

Quinn: "But how do I prevent actually bad consequences from happening?"

Carter: "If your head is full of lies, you can't really trust ad-hoc object-level arguments against speech, like 'saying PADP didn't save very many puppies is bad because PADP is a good charity'. You can instead think about what discourse norms lead to the truth being revealed, and which lead to it being obscured. We've seen, during this

conversation, that appeals to consequences tend to obscure the truth. And so, if we share the goal of reaching the truth together, we can agree not to do those."

Quinn: "That still doesn't answer my question. What about things that are actually bad, like privacy violations?"

Carter: "It does seem plausible that there should be some discourse norms that protect privacy, so that some facts aren't revealed, if such norms have good consequences overall. Perhaps some topics, such as individual people's sex lives, are considered to be banned topics (in at least some spaces), unless the person consents."

Quinn: "Isn't that an appeal to consequences, though?"

Carter: "Not really. Deciding what privacy norms are best requires thinking about consequences. But, once those norms have been decided on, it is no longer necessary to prove that privacy violations are bad during discussions. There's a simple norm to appeal to, which says some things are out of bounds for discussion. And, these exceptions can be made without allowing appeals to consequences in full generality."

Quinn: "Okay, so we still have something like appeals to consequences at the level of norms, but not at the level of individual arguments."

Carter: "Exactly."

Quinn: "Does this mean I have to say a relevant true fact, even if I think it's bad to say it?"

Carter: "No. Those situations happen frequently, and while some radical honesty practitioners try not to suppress any impulse to say something true, this practice is probably a bad idea for a lot of people. So, of course you can evaluate consequences in your head before deciding to say something."

Quinn: "So, in summary: if we're going to have suppression of some facts being said out loud, we should have that through either clear norms designed with consequences (including consequences for epistemology) in mind, or individuals deciding not to say things, but otherwise our norms should be protecting true speech, and outlawing appeals to consequences."

Carter: "Yes, that's exactly right! I'm glad we came to agreement on this."

# Why did we wait so long for the bicycle?

This is a linkpost for https://rootsofprogress.org/why-did-we-wait-so-long-for-the-bicycle

*h/t alyssavance*

The bicycle, as we know it today, was not invented until the late 1800s. Yet it was a simple mechanical invention. It would seem to require no brilliant inventive insight, and certainly no scientific background.

Why, then, wasn't it invented much earlier?

I asked this question on Twitter, and read some discussion on Quora. People proposed many hypotheses, including:

**+ Technology factors.** Metalworking improved a lot in the 1800s: we got improved iron refining and eventually cheap steel, better processes for shaping metal, and ability to make parts like hollow tubes. Wheel technology improved: wire-spoke (aka tension-spoked) wheels replaced heavier designs; vulcanized rubber (1839) was needed for tires; inflatable tires weren't invented until 1887.

Chains, gears, and ball bearings are all crucial parts that require advanced manufacturing techniques for precision and cost.

**+ Design iteration.** Early bicycles were inconvenient and dangerous. The first version didn't even have pedals. Some versions didn't have steering, and could only be turned by leaning. (!) The famous "penny-farthing" design, with its huge front wheel, made it impossible to balance with your feet, was prone to tipping forward on a hard stop, and generally left the rider high in the air, all of which increased risk of injury. It took decades of iteration to get to a successful bicycle model.

**+ Quality of roads.** Roads in the 1800s and earlier were terrible by modern standards. Roads were often dirt, rutted from the passage of many carts, turning muddy in the rain. Macadam paving, which gave smooth surfaces to roads, wasn't invented until about 1820. City roads at the time were paved with cobblestones, which were good for horses but too bumpy for bicycles. (The unevenness was apparently a feature, assisting in the runoff of sewage—leading one Quora answer to claim that the construction of city sewers was what opened the door to bicycles.)

**+ Competition from horses.** Horses were a common and accepted mode of transportation at the time. They could deal with all kinds of roads. They could carry heavy loads. Who then needs a bicycle? In this connection, it has been claimed that the bicycle was invented in response to food shortages due to the "Year without a Summer", an 1816 weather event caused by the volcanic explosion of Mt. Tambora the year earlier, which darkened skies and lowered temperatures in many parts of the world. The agricultural crisis caused horses as well as people to starve, which led to some horses being slaughtered for food, and made the remaining ones more expensive to feed. This could have motivated the search for alternatives.

**+ General economic growth.** Multiple commenters pointed out the need for a middle class to provide demand for such an invention. If all you have are a lot of poor peasants and a few aristocrats (who, by the way, have horses, carriages, and drivers), there isn't much of a market for bicycles. This is more plausible when you realize that bicycles were more of a hobby for entertainment before they became a practical means of transportation.

**+ Cultural factors.** Maybe there was just a general lack of interest in useful mechanical inventions until a certain point in history? But when did this change, and why?

These are all good hypotheses. But some of them start to buckle under pressure:

The quality of roads is relevant, but not really the answer. Bicycles can be ridden on dirt roads or sidewalks (although the latter led to run-ins with pedestrians and made bicycles unpopular among the public at first). And historically, roads didn't improve until *after* bicycles became common—indeed it seems that it was in part the cyclists who called for the improvement of roads.

I don't think horses explain it either. A bicycle, from what I've read, was cheaper to buy than a horse, and it was certainly cheaper to maintain (if nothing else, you don't have to feed a bicycle). And it turns out that inventors were interested in the problem of human-powered vehicles, dispensing with the need for horses, for a long time before the modern bicycle. Even Karl von Drais, who invented the first

> two-wheeled human-powered vehicle after the Year without a Summer, had been working on the problem for years before that.
>
> Technology factors are more convincing to me. They may have been necessary for bicycles to become practical and cheap enough to take off. But they weren’t needed for early experimentation. Frames can be built of wood. Wheels can be rimmed with metal. Gears can be omitted. Chains can be replaced with belts; some early designs even used treadles instead of pedals, and at least one design drove the wheels with levers, as on a steam locomotive.
>
> So what’s the real explanation?

(Continue reading. 2,184 words and lots of great bicycle pictures.)

This post is a single piece from Jason Crawford's project, The Roots of Progress, aptly named, to understand the nature and causes of human progress. I haven't thought deeply enough to check his research, but it's a fascinating project. This essay examines a specific piece of technology, but the case study is used to develop and support models of what it takes for progress to occur.

# How often are new ideas discovered in old papers?

Suppose someone wrote a paper about X two decades ago. A modern reader realizes the X paper sheds light on an unrelated idea Y. Do we have any information on how often this happens? How often is this just "I figured out Y for a different reason, and while doing my lit review I realized that the X paper is also relevant for Y"?

# Let's Read: Superhuman AI for multiplayer poker

On July 11, a new poker AI is published in *Science*. Called Pluribus, it plays 6-player No-limit Texas Hold'em at superhuman level.

In this post, we read through the paper. The level of exposition is between the paper (too serious) and the popular press (too entertaining).

## Basics of Texas Hold'em

If you don't know what it even is, like me, then playing a tutorial would be best. I used Learn Poker on my phone.

Now that you know how to play it, it's time to deal with some of the terminologies.

- Big blind: the minimal money/poker chips that every player must bet in order to play. For example, $0.1 would be a reasonable amount in casual play.
- No-limit: you can bet as much as you want. Okay, not really. You can't bet a billion dollars. In practical playing, it's usually limited to something "reasonable" like 100 times of the big blind.
- Heads-up: 2-player.
- Limping: betting the minimal amount that you have to bet, in order to keep yourself in the game. This is generally considered bad: if you feel confident, you should raise the bet, and if you feel diffident, you should quit.
- Donk betting: some kind of uncommon play that's usually considered dumb (like a donkey). I didn't figure out what it actually means.

## The authors

The authors are Noam Brown and Tuomas Sandholm. Previously, they made the news by writing Libratus, a poker AI that beat human champions in 2-player no-limit Texas Hold'em, in 2017.

Pluribus contains a lot of the code from Libratus and its siblings:

> The authors have ownership interest in Strategic Machine, Inc. and Strategy Robot, Inc. which have exclusively licensed prior game-solving code from Prof. Sandholm's Carnegie Mellon University laboratory, which constitutes the bulk of the code in Pluribus.

Scroll to the bottom for more on the two companies.

## Highlights from the paper

#### Is Nash equilibrium even worthwhile?

In multiplayer games, Nash equilibriums are not easy to compute, and might not even matter. Consider the Lemonade Stand Game:

> It is summer on Lemonade Island, and you need to make some cash. You decide to set up a lemonade stand on the beach (which goes all around the island), as do two others. There are twelve places to set up around the island like the numbers on a clock. Your price is fixed, and all people go to the nearest lemonade stand. The game is repeated. Every night, everyone moves under cover of darkness (simultaneously). There is no cost to move. After 100 days of summer, the game is over. The utility of the repeated game is the sum of the utilities of the single-shot games.

The Nash equilibrium is when three of you are equidistant from each other, but there's no way to achieve that unilaterally. You might decide that you will just stay in Stand 0 and wait for the others to get to Stand 4 and Stand 8, but they might decide upon a different Nash equilibrium.

The authors decided to go all empirical and not consider the problem of Nash equilibrium:

> The shortcomings of Nash equilibria outside of two-player zero-sum games, and the failure of any other game-theoretic solution concept to convincingly overcome them, have raised the question of what the right goal should even be in such games. In the case of six-player poker, we take the viewpoint that our goal should not be a specific game-theoretic solution concept, but rather to create an AI that empirically consistently defeats human opponents, including elite human professionals.

The success of Pluribus appears to vindicate them:

> ... even though the techniques do not have known strong theoretical guarantees on performance outside of the two-player zero-sum setting, they are nevertheless capable of producing superhuman strategies in a wider class of strategic settings.

**Description of Pluribus**

Pluribus first produces a "blueprint" by offline self-play, then during live gaming, adapt it:

> The core of Pluribus's strategy was computed via self play, in which the AI plays against copies of itself, without any data of human or prior AI play used... Pluribus's self play produces a strategy for the entire game offline, which we refer to as the blueprint strategy. Then during actual play against opponents, Pluribus improves upon the blueprint strategy by searching for a better strategy in real time for the situations it finds itself in during the game.

Since the first round (like chess opening vs chess midgame) had the smallest amount of variation, Pluribus could afford to train an almost complete blueprint strategy for the first round. For later rounds, some real-time search was needed:

> Pluribus only plays according to this blueprint strategy in the first betting round (of four)... After the first round, Pluribus instead conducts real-time search to determine a better, finer-grained strategy for the current situation it is in.

Pluribus uses Monte Carlo counterfactual regret minimization. The details can be found in the link.

> The blueprint strategy in Pluribus was computed using a variant of counterfactual regret minimization (CFR)... We use a form of Monte Carlo CFR (MCCFR) that samples actions in the game tree rather than traversing the entire game tree on each iteration.

Pluribus can be sneaky:

> ... if the player bets in [a winning] situation only when holding the best possible hand, then the opponents would know to always fold in response. To cope with this, Pluribus keeps track of the probability it would have reached the current situation with each possible hand according to its strategy. Regardless of which hand Pluribus is actually holding, it will first calculate how it would act with every possible hand, being careful to balance its strategy across all the hands so as to remain unpredictable to the opponent. Once this balanced strategy across all hands is computed, Pluribus then executes an action for the hand it is actually holding.

This was corroborated by a comment from a human opponent:

> "Pluribus is a very hard opponent to play against," said Chris Ferguson, a World Series of Poker champion. "It's really hard to pin him down on any kind of hand."

Scroll down for how Ferguson lost to Pluribus.

**Pluribus is cheap, small, and fast**

In order to make Pluribus small, the blueprint strategy is "abstracted", that is, it intentionally confuses some game actions (because really, $200 and $201 are not so different).

> We set the size of the blueprint strategy abstraction to allow Pluribus to run during live play on a machine with no more than 128 GB of memory while storing a compressed form of the blueprint strategy in memory.

The abstraction paid off. Pluribus was cheap to train, cheap to run, and faster than humans:

> The blueprint strategy for Pluribus was computed in 8 days on a 64-core server for a total of 12,400 CPU core hours. It required less than 512 GB of memory. At current cloud computing spot instance rates, this would cost about $144 to produce.
>
> When playing, Pluribus runs on two Intel Haswell E5-2695 v3 CPUs and uses less than 128 GB of memory. For comparison... Libratus used 100 CPUs in its 2017 matches against top professionals in two-player poker.

On Amazon right now, Intel® Xeon® Processor E5-2695 v3 CPU cost just $500 each, and a 128 GB RAM cost $750. The whole setup can be constructed for under $2000. It would only take a little while to recoup the cost if it goes to online poker.

> The amount of time Pluribus takes to conduct search on a single subgame varies between 1 s and 33 s depending on the particular situation. On average, Pluribus

> plays at a rate of **20 s per hand when playing against copies of itself** in six-player poker. This is roughly **twice as fast as professional humans tend to play**.

**Pluribus vs Human professionals. Pluribus wins!**

> We evaluated Pluribus against elite human professionals in two formats: five human professionals playing with one copy of Pluribus (5H+1AI), and one human professional playing with five copies of Pluribus (1H+5AI). Each human participant has won more than $1 million playing poker professionally.

Professional Poker is an endurance game, like marathon:

> In this experiment, 10,000 hands of poker were played over 12 days. Each day, five volunteers from the pool of [13] professionals were selected to participate based on availability. The participants were not told who else was participating in the experiment. Instead, each participant was assigned an alias that remained constant throughout the experiment. The alias of each player in each game was known, so that players could track the tendencies of each player throughout the experiment.

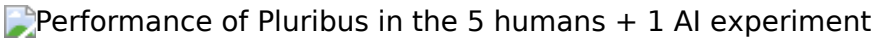
And there was prize money, of course, for the humans. Pluribus played for free -- what a champ.

> $50,000 was divided among the human participants based on their performance to incentivize them to play their best. Each player was guaranteed a minimum of $0.40 per hand for participating, but this could increase to as much as $1.60 per hand based on performance.

Pluribus had a very high win rate, and is statistically demonstrated to be profitable when playing against 5 elite humans:

> After applying AIVAT, Pluribus won an average of 48 mbb/game (with a standard error of 25 mbb/game). This is considered a very high win rate in six-player no-limit Texas hold’em poker, especially against a collection of elite professionals, and implies that Pluribus is stronger than the human opponents. Pluribus was determined to be profitable with a p-value of 0.028.

"mbb/game" means "milli big blinds per game". "big blind" just means "the least amount that one must bet at the beginning of the game", and poker players use it as a unit of measurement of the size of bets. "milli" means 1/1000. So Pluribus would on average win 4.8% of the big blind each game. Very impressive.

 Performance of Pluribus in the 5 humans + 1 AI experiment

AIVAT is statistical technique that is designed specifically to evaluate how good a poker player is. From (Neil Burch et al, 2018):

> Evaluating agent performance when outcomes are stochastic and agents use randomized strategies can be challenging when there is limited data available... [AIVAT] was able to reduce the standard deviation of a Texas hold’em poker man-machine match by 85% and consequently requires 44 times fewer games to draw the same statistical conclusion. AIVAT enabled the first statistically significant AI victory against professional poker players in no-limit hold’em.

**Pluribus vs Jesus (and Elias)**

> The human participants in the 1H+5AI experiment were Chris “Jesus” Ferguson and Darren Elias. Each of the two humans separately played 5,000 hands of poker against five copies of Pluribus.

Pluribus did not gang up on the poor human:

> Pluribus does not adapt its strategy to its opponents and does not know the identity of its opponents, so the copies of Pluribus could not intentionally collude against the human player.

The humans were paid on average $0.60 per game:

> To incentivize strong play, we offered each human $2,000 for participation and an additional $2,000 if he performed better against the AI than the other human player did.

Pluribus won!

> For the 10,000 hands played, Pluribus beat the humans by an average of 32 mbb/game (with a standard error of 15 mbb/game). Pluribus was determined to be profitable with a p-value of 0.014.

Ferguson lost less than Elias:

> Ferguson’s lower loss rate may be a consequence of variance, skill, and/or the fact that he used a more conservative strategy that was biased toward folding in unfamiliar difficult situations.

**Pluribus is an alien, like AlphaZero**

And like AlphaZero, it confirms some human strategies, and dismisses some others:

> Because Pluribus’s strategy was determined entirely from self-play without any human data, it also provides an outside perspective on what optimal play should look like in multiplayer no-limit Texas hold’em.

Two examples in particular:

> Pluribus confirms the conventional human wisdom that limping (calling the “big blind” rather than folding or raising) is suboptimal for any player except the “small blind” player... While Pluribus initially experimented with limping... it gradually discarded this action from its strategy as self play continued. However, Pluribus disagrees with the folk wisdom that “donk betting” (starting a round by betting when one ended the previous betting round with a call) is a mistake; Pluribus does this far more often than professional humans do.

**Too dangerous to be released, again**

The program is not released for some kind of unspecified risk. (News articles made it specifically about the risk of wrecking the online gambling industry.)

> Because poker is played commercially, the risk associated with releasing the code outweighs the benefits. To aid reproducibility, we have included the pseudocode for the major components of our program in the supplementary materials.

# Useful quotes from other news report

From Ars Technica:

> Pluribus actually confirmed one bit of conventional poker-playing wisdom: it's just not a good idea to "limp" into a hand, that is, calling the big blind rather than folding or raising. The exception, of course, is if you're in the small blind, when mere calling costs you half as much as the other players.
>
> Pluribus placed donk bets far more often than its human opponents... Pluribus makes unusual bet sizes and is better at randomization. "Its major strength is its ability to use mixed strategies... to do this in a perfectly random way and to do so consistently. Most people just can't."

From MIT Technology Review:

> Sandholm cites multi-party negotiation or pricing—such as Amazon, Walmart, and Target trying to come up with the most competitive pricing against each other—as a specific application. Optimal media spending for political campaigns is another example, as well as auction bidding strategies.

There are a bit of details to the two companies of Sandholm:

> Sandholm has already licensed much of the poker technology developed in his lab to two startups: Strategic Machine and Strategy Robot. The first startup is interested in gaming and other entertainment applications; Strategy Robot's focus is on defense and intelligence applications.

"Better computer games"... hm, sounds suspiciously nonspecific.

> Brown says Facebook has no plans to apply the techniques developed for six-player poker, although they could be used to develop better computer games.

# How to take smart notes (Ahrens, 2017)

This is my rephrasing of (Ahrens, 2017, *How to Take Smart Notes*). I added some personal comments.

## The amazing note-taking method of Luhmann

To be more productive, it's necessary to have a good system and workflow. The Getting Things Done system (collect everything that needs to be taken care of in one place and process it in a standardised way) doesn't work well for academic thinking and writing, because GTD requires clearly defined objectives, whereas in doing science and creative work, the objective is unclear until you've actually got there. It'd be pretty hard to "innovate on demand". Something that can be done on demand, in a predetermined schedule, must be uncreative.

Enter Niklas Luhmann. He was an insanely productive sociologist who did his work using the method of "slip-box" (in German, "Zettelkasten").

Making a slip-box is very simple, with many benefits. The slip-box will become a research partner who could "converse" with you, surprise you, lead you down surprising lines of thoughts. It would nudge you to (number in parenthesis denote the section in the book that talks about the item):

- Find dissenting views (10.2, 12.3)
- Really understand what you learned (10.4, 11.2, 11.3, 12.6)
- Think across contexts (12.5)
- Remember what you learned (11.3, 12.4)
- Be creative (12.5, 12.6, 12.7, 13.2)
- Get the gist, not stuck on details (12.6)
- Be motivated (13.3)
- Implement short feedback loops, which allows rapid improvements (12.6, 13.5)

## Four kinds of notes

### Fleeting notes

These are purely for remembering your thoughts. They can be: fleeting ideas, notes you would have written in the margin of a book, quotes you would have underlined in a book.

They have no value except as stepping stones towards making literature and permanent notes. They should be thrown away as soon as their contents have been transferred to literature/permanent notes (if worthy) or not (if unworthy).

Examples:

> Jellyfish might be ethically vegan, since they have such a simple neural system, they probably can't feel pain.

> Ch. 9 How to ~~attention~~ attend:
>
> 1. One thing at a time. No multitasking
> 2. When writing, attend to idea flow. Meaning, not wording. ...

**Literature notes**

These summarize the content of some text, and give the citation.

Example:

> (Kahneman & Tversky, 1973) shows that people often do not take into account the prior when doing a Bayesian probability problem. In particular, when no evidence is given, the prior probabilities are used; when worthless evidence is given, prior probabilities are ignored.

---

> Kahneman, Daniel, and Amos Tversky. “On the Psychology of Prediction.” Psychological Review (1973)

Such notes could be made in Zotero, which is how I do it. You might make them separately in some other notebook software, or just in plain text files.

**Permanent notes**

Each permanent note contains one idea, explained fully, in complete sentences, as if part of a published paper.

There are many tools available for storing the permanent notes, see Tools • Zettelkasten Method. I personally recommend TiddlyWiki.

**Project notes**

These are notes made only for a project, such as a note that collects all the notes that you'd want to assemble into a paper. They can be thrown away after the project is finished.

# Four principles

**Writing is the only thing that matters.**

Don't just read. Make reading notes. Don't just learn. Make blog posts or something to share what you learned.

Also, hand-written notes has some advantage. In (Mueller & Oppenheimer, *The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking*, 2014), it's shown that students who take notes by laptop understood lectures less, due to their tendency to transcribe verbatim without understanding. From mouth to ears to fingers, bypassing the brains completely.

The way I see it, this is not an argument against using the computer, but an argument for repharsing instead of copy-pasting/direct quoting/mere transcribing.

### Be simple

Don't underline, highlight, write in the margins, or use several complicated systems for annotation. It'd make it really hard for you to retrieve these scattered ideas later. You would be forced to remember with your biological brain to keep track of what information is put where.

Put all these ideas in the same simple system of your slip-box, and you will be set free to use your biological brain to think about these ideas.

Your simple slip-box system would be like an external brain that interfaces seamlessly with your biological brain.

### Papers are linear, but writing is nonlinear

This is why advice on "how to write" in the form of a list of "do this then that" is bound to do badly.

Instead, you should write a lot of permanent notes in your slip-box. Then when the time comes for you to write a paper, just select a linear path out of the network of notes, then rephrase and polish that into a paper.

Calculate productivity not by how many pages of paper you've written, but by how many permanent notes you've written per day. This is because some pages of a paper can take months to write, others can take hours. In contrast, each permanent note takes roughly the same amount of time to write.

### Short feedback loops

Feedback loops should be short. It makes you learn fast, fail fast, succeed fast. According to (Kahneman & Klein, *Conditions for Intuitive Expertise: A Failure to Disagree*, 2009), this is how intuitive expertise is made: a lot of practice in an environment with rapid and unambiguous feedback.

The traditional way of writing a paper takes months before you get a feedback in the form of reviewers' comments. Instead, you should make notes, which you could make several per day, allowing fast feedback loops. If you really understood something, you'd see it in the form of a well-written note. If not, then you know you haven't really understood it. You can experiment with other ways to make the notes and you will see immediately what works and what doesn't.

# Six methods

### How to pay attention

Don't multitask. Pay attention to one task at a time.

When writing, pay attention to the idea flow, what you want the words to mean. Don't pay attention to what the words actually mean.

When proofreading, pay attention to what the words are saying, and not what you think they mean.

Pay attention only to what you must and don't pay attention to anything else, because attention is very precious.

Routinize things that can be routinized, such as food, water, clothes... Wear only one outfit ever, like Steve Jobs. Eat only one meal plan, buy exactly the same kind of groceries, or better, always eat the first vegan meal plan at the canteen.

Use the Zeigarnik effect to your advantage. If you want something to stop intruding your mind, write it down and promise yourself that you'll "deal with it later". If you want to keep pondering something (perhaps a problem you want to solve), don't write it down, and go for a walk with that problem on your mind.

### How to make literature notes

As mentioned before, each literature note contains exactly two parts: the content of a text, and the bibliographical location of the text. If you do the note in a bibliography software like Zotero, you can attach the note directly to the text, and there's no need for the bibliography information.

The most important thing is to capture **your** understanding of the text, so don't quote. Quoting can easily lead to out-of-context quoting. Preseve the context as much as possible by paraphrasing.

Prepare the literature notes so that when you make permanent notes, you can elaborate on the texts, that is, describe the context, find connections and contrasts and contradictions with other texts.

### How to make permanent notes

Recontexutalize ideas in **your** thought. Write down why you would care about an idea. For example (from section 11.2), if the idea is an observation from (Mullainathan and Shafir, 2013, *Scarcity: Why having too little means so much*):

> people with almost no time or money sometimes do things that don’t seem to make any sense... People facing deadlines sometimes switch frantically between all kinds of tasks. People with little money sometimes spend it on seeming luxuries like take-away food.

Then

> As someone with a sociological perspective on political questions and an interest in the project of a theory of society, my first note reads plainly:
>
> > Any comprehensive analysis of social inequality must include the cognitive effects of scarcity. Cf. Mullainathan and Shafir 2013.

### How to link between notes

There are three kinds of links between notes:

- Index -> Entry point note
- Note -> Note
- Note <-> Note

At the top level, there is one note called "Index". The index note is just a list of tags/keywords with links. Each tag/keyword is a topic that **you** care about, and is linked to a few notes (Luhmann limited himself to at most 2) that serve as "entry points" to the topic.

The entry points are often notes that give overviews to the topic. Luhmann would make these notes to be an annotated list of notes that cover various aspects of the topic. His entry-point notes would have list length up to 25.

Between notes, there are two kinds of links: sequential and horizontal. In fact, sequential links are really just horizontal links that you annotate as "sequential".

For example, consider this note:

> Following: [link 1] [link 2]...
>
> ---
>
> Content content [link 3] content content [link 4]...
>
> ---
>
> Followed by: [link 5] [link 6] ...

After reading this note, you can go along the sequence and read "Followed by" notes, or take a sideways stride and follow the horizontal link [link 3].

The advantage of marking some links as sequential is that you get clear sequences of thought that you can easily follow, but they are by no means essential. You could just make horizontal links.

Ideally, you should make the network of slip-box notes to be like a small-world network, with a few notes having many connections, and some notes having "weak ties" to far-away notes (Granovetter, Mark S, 1977 *The Strength of Weak Ties*).

**How to write a paper**

Don't brainstorm, since brainstormed ideas are what's easily available, instead of innovative or actually relevant. Especially don't group-brainstorm, which tend to become even less innovative due to groupthink effects (Mullen, Brian, Craig Johnson, and Eduardo Salas, 1991, *Productivity Loss in Brainstorming Groups: A Meta-Analytic Integration*).

Instead, do a walk through the slip-box and select a linear path. That gives you a draft from which you can polish into a paper.

Work on several papers simultaneously, switch if bored. This is a kind of "slow multitasking", which is good multitasking. Luhmann said

> When I am stuck for one moment, I leave it and do something else... I always work on different manuscripts at the same time. With this method, to work on different things simultaneously, I never encounter any mental blockages.

When you need to cut out something that you really like, but just doesn't belong to the paper (such as something that is not relevant to the argument), you can make a file named "maybe later.txt" and dump all the things that you promise to add back later (but never actually do). This is a psychological trick that works.

## How to start the habit of using slip-boxes

Old habits die hard. The best way to break an old habit is to make a new habit that can hopefully replace the old habit.

For getting into the habit of using slip-boxes, you can start by making literature notes. Once you have that habit, making permanent notes would be a natural next habit to take on.

# Are we certain that gpt-2 and similar algorithms are not self-aware?

Have someone even started this conversation? This is f*cked up. I'm really, really freaked out lately with some of those.
https://old.reddit.com/r/SubSimulatorGPT2/comments/cbauf3/i_am_an_ai/

And I read up a lot about cognition and AI, and I'm pretty certain that they are not. But shouldn't we give it a lot more consideration just because there's a small chance they are? Because the ramifications are that dire.

But then again, I'm not sure that knowledge will help us in any way.

# What are good resources for learning functional programming?

I'm looking to use the 3 Books Technique to learn functional programming. Does anyone have any "What", "Why' or "How" resources for functional programming (or resources that don't fit the categories?)

# Intellectual Dark Matter

**Knowledge that we can show exists, but cannot directly access, rests at the foundations of society and technology.**

*Messier 106*

*From my post published on the Long Now Foundation's blog. This is an excerpt from the draft of my upcoming book on great founder theory.*

## Missing mass, missing knowledge

Many galaxies would fly apart if they had as much mass as estimates based on their visible signature suggest. Although some have posited alternative theories of gravitation to explain this discrepancy, most physicists now hypothesize the existence of mass-bearing particles that are not detectable through emitted radiation such as visible light. We call these particles dark matter, and it is estimated to compose about 85% of all matter in the observable universe.

In analyzing the functional institutions of our society, we are not able to see for ourselves most of the knowledge that created them. Knowledge of this sort includes trade secrets, tacit technical knowledge, private social networks, private intelligence-gathering operations, management and persuasive skill, cooperation and collusion among founders and their allies, and founders' long-term plans for their institutions.[1]

This knowledge has profound effects on the social landscape. We must understand it if we hope to understand society. We therefore must examine **intellectual dark matter**: knowledge we cannot see publicly, but whose existence we can infer because our institutions would fly apart if the knowledge we see were all there was.[3] Such intellectual dark matter rests at the foundations of our society, dwarfing in scope and importance the accessible, shareable, visible knowledge on which we normally focus.

There are many forms of intellectual dark matter, but the three principal ones are lost, proprietary, and tacit knowledge.

*Read the rest here.*

*Read more from Samo Burja here.*

# Where is the Meaning?

"You're being mean."
"No, I'm not."

"This album is amazing!"
"What are you talking about? It's clearly trash."

"Tomatoes are my favorite vegetable!"
"Dude, they're fruits."

People often argue about what things *are*, and they often do it with words. Sometimes this can create another argument about what a certain word *means*. This leads me to semi-rhetorically ask you the question, where is the meaning located? In the words? The people? The air? If this seems like philosophical thumb-twiddling, note that where you think meaning is determines where you go looking whenever a question of meaning pops up. If I think the dictionary is the end all be all of meaning, it's the first place I'll go when pondering, "What really is happiness?"

If you've ever been quoted out of context in a way that made you look bad, you might have grumbled, "That's not what I meant!" On some level, I think most people understand that words themselves do not have inherent meaning. There are thoughts in your head, and words are the things that come out of your mouth, which hopefully let others guess what you are thinking.

And yet... damn sometimes it just really *feels* like meaning is baked into a message. Like it's just there, waiting to be discovered. We're going to explore this feeling, what I think it means, and how to not let it trip you up.

# Context and Predictability

Context matters:

"My sister just got a new dog."
vs
"You snuck in? You sly dog!"

Context still matters:

Mobster: "Watch do to that guy who owed you money?"
Mob Boss: "I put him to sleep."
vs
Spouse 1: "Where's our Henry? I haven't seen him in a bit"
Spouse 2: "I put him to sleep."

Neither of those are meant to be shocking, just reminders that you already know that the same word, even the same sentence, can mean completely different things based on the context.

Everyone has the shared context of "having a human brain" and "being subject to the same laws of physics". This lets me communicate things like "Get away! You're not welcome" merely by throwing rocks. Language is a wild invention that was bootstrapped from an incredibly general context, and now it lets us create even more shared context which allows for even more specific thoughts and ideas to be shared. The context that you share with the people you're talking to is incredibly important in figuring out who meant what.

Given that it seems pretty clear that context is essential to understanding the meaning of someone's words, let's hop back to examine the feeling that meaning is intrinsic to words. Douglas Hofstadter talks about this feeling in Godel, Escher, Bach (the chapter "The Location of Meaning" really digs into some interesting ideas about meaning that are beyond the scope of this post). In the context of the Rosetta Stone being translated, he says:

> Just how intrinsic is the meaning of a text, when such mammoth efforts are required in order to find the decoding rules? Has one put meaning into the text, or was that meaning already there? My intuition says that the meaning was always there, and that despite the arduousness of the pulling-out process, no meaning was pulled out that wasn't in the text to start with. This intuition comes mainly from one fact: I feel that **the result was inevitable;** that, had the text not been deciphered by this group at this time, it would have been deciphered by that group at that time-and it would have out the same way. That is why the meaning is part of the text itself; it acts upon intelligence in a predictable way. Generally, we can say: meaning is part of an object to the extent that it acts upon **intelligence in a predictable way.**
> (emphasis mine)

Predictability is key. You can see how if saying "Good morning, what's the time?" prompted some people to give you a handshake, others to tell you their cell phone number, and others to start dancing, you'd be far less inclined to think that "Good morning, what's the time?" had intrinsic meaning.

I claim that "It feels like meaning is intrinsic to the words themselves" is a thought composed of multiple steps:

1. I notice that some words predictably convey a particular meaning.
2. When I communicate, say in writing, I'm only giving the other person words.
3. If I'm only giving them some words, and they reliable get the same meaning, it must be because the meaning is *some how* baked into the words.

And now that we've drawn out the implicit jumps of logic hidden, we can see that step 2 is a bit fishy. Yes, you might have only given someone your words, but what they already had was some context. Context from having grown up in the same country as you, context from having know you from work, context from earlier parts of your conversation. When a context is very familiar you can fuse to it. When this happens you no longer notice your context when you look at the world, because the context becomes the **lens through which you do your looking**. If you turn your attention to these mysterious words that seem to so predictably convey your meaning, they are the only culprits available and you quite reasonably decide that *they* must contain the meaning.

There is a right answer to "What did you mean by [words you said]?" but there is not a right answer to "What do [words you said] mean?"

# Go to the source of the meaning

A sketch:

> *Dale is chopping up a tomato and putting it into a fruit salad*
>
> Brice: What are you doing?! A tomato is a vegetable, don't put it in a fruit salad!
>
> Dale: Nope, I looked it up, totally a fruit.
>
> *Dale and Brice spend 10 minutes having an unproductive argument about whether a tomato is a fruit or a vegetable, it ends with them agreeing to disagree*

Brice said some words. What does the word fruit mean? Let's investigate the definition of fruit.

Here's a parallel universe where Brice has read A Hazardous Guide To Words (or the Sequences, but then again Brice is reading HGTW because he can't be bothered to read the Sequences):

> *Dale is chopping up a tomato and putting it into a fruit salad*
>
> Brice: What are you doing?! A tomato is a vegetable, don't put it in a fruit salad!
>
> Dale: Nope, I looked it up, totally a fruit.
>
> Brice: Sorry, what I meant was that the flavors of those bananas, strawberries, blueberries, and orange that you already have in the fruit salad won't really jive with the flavor of the tomato, and I think you shouldn't include it.
>
> *Dale and Brice spend 10 minutes having an interesting conversation about what sorts of flavors do and don't go well together. It ends with them both learning something.*

Brice said some words. What did Brice mean? Let's have Brice tell us more.

Remember, words don't mean things, **people mean things**. And because language is awesome, plenty of words are particularly good indicators that a person meant a particular thing.

# Is there neuroscience research on cognitive biases?

I recently watched NeuraLink's presentation, and wondered how can something like that help us reason.

The obvious way is an AI that can reason, and is connected to our brains and we reason together with it.

But another direction i thought of, is just helping us notice when we're using motivated cognition and letting cognitive biases take place.

Another thing i thought of was reducing Akrasia, can it help us win the fight between areas in the brain in an akratic situation?

With my very little knowledge about the subjects in hand, it seems like it would be an easier target - seems to me you'd need fewer electrodes, less understanding of how reasoning works, and simpler software.

Although the questions about the possibilities of this technology are interesting, i don't expect anything more the guesses and predictions to be available right now.

So my question is whether we have neurological knowledge about how these mechanisms work.

# Prediction as coordination

I want to introduce a model of why forecasting might be useful which I think is underappreciated: it might help us solve coordination problems.

This is currently only a rough idea, and I will proceed by examples, pushing this post out early rather than not at all.

### The standard model of forecasting

This looks something like:

> We have our big, confusing, philosophical, long-term uncertainties. We then need to 1) find the right short-term questions which capture these uncertainties, which are 2) understandable to traditional Superforecasters without very deep inside knowledge, and whose expertise has only been demonstrated on short-term questions in more well-understood domains, who then 3) use tools like outside views and guesstimates to answer them.

When I hear people say they're not excited about forecasting, it's almost always because they think this standard model won't work for AI safety. I'm very sympathetic to that view.

### Example 1: coordination in mathematics via formalism

When quantifying our beliefs, we lose a large amount of nuance and interpretability. This is similar to how, when formalising things mathematically, we sacrifice the majority of human understanding.

What we gain, instead, is the ability to express and communicate thoughts...

- much more succinctly
- using a precise, interpersonally standardised format of interpretation
- in a way which clarifies certain logical and conceptual relations

This is a trade-off that allows a *community* of mathematicians to make intellectual progress together, and to effectively make results common knowledge in a way which allows them to coordinate on what to solve next.

### Example 2: futures markets as using predictions for coordination

Getting enough food for everyone is a big coordination problem. We want some people to stockpile things like rice and wheat so that we're prepared for a drought, but we also don't want people to waste opportunities on storing stuff which has to be thrown away in case the next harvest goes well. These kinds of problems are solved by futures markets, which effectively predict the future price of rice, and thereby provide an incentive arbitrage away any abrupt price fluctuations (i.e. to strategically

stockpile/sell out rice so as to match future supply and demand). Robert Shiller has suggested these as a candidate for the greatest financial innovation.

**Example 3: predicting community consensus**

One particularly interesting use case is trying to predict what the x-risk community will believe at some time *t* in the future. Assuming the community is truth-seeking, anyone who spots the direction in which opinions will converge *in advance* of their convergence has 1) performed an important epistemic service, and 2) provided important evidence of their own epistemic trustworthiness.

For example, the CAIS model has gathered a fair amount of attention. (I personally don't have a strong inside view on it.) If someone would have predicted this shift more than a year ago, we would want to trust them a bit more next time they predicted a shift in community attention.

It was mentioned to me that one researcher thought this model important more than 1.5 years ago; but the reason he thought so was not because of superior reasoning -- but because of inside knowledge.

This is an inefficiency. The frontiers of our collective attention allocation do not line up with the frontiers of our intellectual progress, and hence see abrupt fluctuations as papers are released and the advantage of inside info is dispelled.

One implementation of this might look like sending out a survey asking about important questions to important organisations on ~yearly intervals, and then have people trying to predict the outcomes of that survey.

This has one important advantage over the standard uses of forecasting: we don't have to resolve the questions "all the way down". If we simply ask what people will *think* take-off speeds are likely to be, rather than what take-off speeds are *actually* going to be, and further assume that people move closer to the truth in expectation, this gives us a much cheaper signal to evaluate.

**Example 4: avoiding info-cascades**

Info-cascades occur when people update off of each others beliefs without appropriately sharing the evidence for those beliefs, and the same pieces of evidence end up "double-counted". Having better system for tracking who believes what and why could help solve this, and prediction systems could be one way of doing so.

**Example 5: building fire alarms**

Eliezer notes that rather than being evidence of a fire, fire alarms make it common knowledge that it's social acceptable to act *as if* there's a fire. They're the cue on which everyone jumps from one social equilibrium to another.

Eliezer claims there's no fire alarm for AGI in society more broadly. I suspect there are also areas within the x-risk space where we don't have fire alarms. Prediction systems

are one way of building them.

### Who is going to make the forecasts?

An important clarification: I'm *not* saying that we should "outsource" the intellectual work of solving hard x-risk research problems to forecasters without domain-expertise. (That is *another* interesting and controversial proposal one might discuss.)

Rather, I'm saying that we should use predictions as a *vehicle* to capture the changing beliefs of *current* domain-experts, and allocate their attention going forwards, (smoothing out attentional discontinuities in expectation).

I'm not saying we should replace Eric Drexler with a swarm of hobby forecasters. I'm saying that a few full-time x-risk researchers might realise before the rest of the community that Eric's work deserves marginally more attention, *and be right about that*, and that community-internal forecasting systems can allow us to more effectively use their insights.

### Can't we just use blog posts?

Compared to a numerical prediction, a blog post...

- ...takes more effort to produce
- ...might take more effort to read and interpret
- ...doesn't have a standardized format of interpretation
- ...doesn't allow gathering and visualisation of the beliefs of multiple people
- ...doesn't natively update in the light of new information

Blog posts have the crucial property of being "essay complete" in their expressiveness, but that comes at the cost of idiosyncracy and poor scalability.

A better model is probably to treat blog posts as part of the ground truth over which predictions operate, just as the rice and wheat markets provide the ground truth for their respective futures markets.

I'd rather have only blog posts than only prediction systems, but I'd rather have *both* than only blog posts.

# Commentary On "The Abolition of Man"

C.S. Lewis wrote a short book attacking moral subjectivism in education; it's available online here as a pdf, here as raw text, and here as a series of videos [1 2 3], and I think probably worth reading in full (at 50 pages or ~100 minutes of video at 1x speed). This post is mostly me rambling about what I saw as the central point, especially connected to individual development and community health, by quoting sections and then reacting to them.

The book begins with a reaction to a grammar textbook (written in 1939) euphemistically called *The Green Book* whose lessons are also philosophical; Lewis doesn't object to the bait-and-switch (outside of one paragraph) so much as the content and quality of the philosophy. (One suspects Lewis wouldn't object to the Copybook Headings referenced by Kipling, even tho that mixes writing lessons and philosophy.)

> Until quite modern times all teachers and even all men believed the universe to be such that certain emotional reactions on our part could be either congruous or incongruous to it--believed, in fact, that objects did not merely receive, but could *merit*, our approval or disapproval, our reverence or our contempt.

First, let's get the obvious objections out of the way: the claim of universality is probably false. Even supposing it were true, then the underlying change seems worth investigating. Naive belief that one's map is objective reality disintegrates on contact with different maps and after noticing surprising divergences between one's predictions and observations; one can imagine this happening in the moral realm as well as the physical one. But presumably we should just ignore this as standard "the contemporary world is fallen and bad" framing instead of an actual historical claim.

The more interesting claim here is the question of whether or not there can or should be a question of *merit*, distinct from a question of flavor or fact. A previous taxonomy I've liked a lot (that I was mostly introduced to by Sapiens) is the split between objective (determined by reality), subjective (determined by the person in question), and intersubjective (determined by some group process); the rules of a game are not just 'my personal whims' and are also not 'scientific' in the sense that any outside observer would be able to determine it themselves. Without access to human civilization; aliens would figure out the same physics, and they might play something like chess, but they likely won't play chess. Nevertheless, concepts like chess are an important component of your epistemology and there is such a thing as a 'legal move' or 'illegal move.'

> But what is common to [religious traditions] is something we cannot neglect. It is the doctrine of objective value, the belief that certain attitudes are really true, and other really false, to the kind of thing the universe is and the kind of things we are. Those who know the *Tao* can hold that to call children delightful or old men venerable is not simply to record a psychological fact about our own parental or filial emotions at the moment, but to recognize a quality which *demands* a certain response from us whether we make it or not.”

Lewis is trying to go a step further; in my framing, there's a thing about the 'game that is society' that involves 'playing with reality' in a way that makes it something a little more objective than the 'intersubjective.' It's not just that everyone jointly decided that old people are venerable and thus the fashion is to venerate them; it's that somehow venerating old people is congruous with the Tao and not venerating them isn't, and so getting that question wrong is worse on some dimension than just playing chess by the wrong rules. Play chess by the wrong rules and people will throw you out of the chess club; play society by the wrong rules and your society collapses or misery abounds. Lewis uses 'the *Tao*' to refer to both 'the underlying territory as distinct from the map' and 'the sort of human behavior congruous with the territory', in a way that seems connected to this sense of 'the universe as participant in the game that is society.'

Note that he says "true to the kind of thing the universe is and the kind of things we are", as opposed to simply "true." This seems consistent with 'morality as the product of game theory', and a sort of subjectivism that allows for different environments to have different moralities, or different professions to have different ethics; the *Tao* of the soldier may be distinct from the *Tao* of the doctor, and the *Tao* of the Inuit different from the *Tao* of the Swahili. It reminds me of the claim that Probability is Subjectively Objective; if one is a soldier, the 'right way to be' is different than if one is a doctor, but there is still a meaningful sense in which there is only 'one right way to be' that is not destroyed by that variation. [Imagine a function from 'broad situation' to 'proper behavior'; this function can vary as you change the input while still being a deterministic function.]

> If they embark on this course the difference between the old and the new education will be an important one. Where the old initiated, the new merely ‘conditions’. The old dealt with its pupils as grown birds deal with young birds when they teach them to fly; the new deals with them more as the poultry-keeper deals with young birds--making them thus or thus for purposes of which the birds know nothing. In a word, the old was a kind of propagation--men transmitting manhood to men; the new is merely propaganda.

The contrast between 'initiation' and 'conditioning' stuck out to me. One way you could get such a split is a separation between Educators and Students where most students will not become educators, whereas most boy-children become men. When I try to figure out what the difference between religions and cults are, especially when it comes to things like the rationality community, I keep thinking about this sense of "explorers trying to create more explorers", and how it differs from "carnies trying to use marks", and somehow it seems connected to feedback loops. The man trying to make the next generation into men relates to the next generation differently from how the carnie trying to extract money from marks relates to those marks. Not only does the former involve identification with the audience (where the latter recoils from that), the former is trying to get the audience to understand the whole process (so that they too, in their time, can perform it), whereas the latter is trying to get the audience to misunderstand the whole process (so that they will come back and be fleeced again).

To the extent that the High Modernist or Reformer or Rationalist sees the outside as a thing to be optimized, as opposed to part of a system that needs to support further optimization, it seems like there's some deep short-sightedness and disconnection from the *Tao*. To the extent that some profession sees the outside world as something to be profited from, as opposed to a body in which they are an organ, we should expect the society to be sick in some way.

> Let us suppose for a moment that the harder virtues could really be theoretically justified with no appeal to objective value. It still remains true that no justification of virtue will enable a man to be virtuous. Without the aid of trained emotions the intellect is powerless against the animal organism. ... The head rules the belly through the chest--the seat, as Alanus tells us, of Magnanimity, of emotions organized by trained habit into stable sentiments. The Chest-Magnanimity-Sentiment--these are the indispensable liaison officers between cerebral man and visceral man. It may even be said that it is by this middle element that man is man; for by his intellect he is a mere spirit and by his appetite mere animal.
>
> The operation of *The Green Book* and its kind is to produce what may be called Men without Chests. It is an outrage that they should be commonly spoken of as Intellectuals. This gives them the chance to say that he who attacks them attacks Intelligence.

This reminded me of Bayesians vs. Barbarians, with a new dimension added; it is not that the Barbarians gain from having less in their head, it is that the Bayesians lost because they forgot to develop their chests. When I was younger, I read through The Fountainhead and Atlas Shrugged and was confused by the educational strategy; here were these staunchly moral characters, as evidenced by their disgust at taking immoral actions that would benefit them, but the source of their morality seemed unspoken and unjustified. This felt like a serious contrast to what I observed at my local church, where people put in serious amounts of effort to become slightly more aligned with their reasoned values. It looked like all that was assumed unnecessary; one simply had to paint the picture of correctness and it would be followed by the righteous without any exercise or training.

Another Eliezer reference is Feeling Rational, which points at the congruity property of emotions, but only with regards to factual truth; if you're afraid about an iron being hot and it's cold, you're making a mistake, and if you're calm about an iron being cold and it's hot, you're making a mistake. But that seems to miss the intersubjective angle; in some contexts, reacting to criticism with defensiveness is inappropriate and reacting to criticism with curiosity is appropriate, and some large part of 'training human rationality' is inculcating the right emotional responses in oneself. A dojo isn't just about transfer of technique, but also about transfer of attitude.

# Do bond yield curve inversions really indicate there is likely to be a recession?

I've seen some recent commentary that the bond yield curve indicates there is likely to be a recession in the next year.

For example, from the New York fed:
https://www.newyorkfed.org/.../r.../capital_markets/Prob_Rec.pdf

and commentary:

https://www.newyorkfed.org/medialibrary/media/research/capital_markets/Prob_Rec.pdf
https://www.wsj.com/.../government-bond-market-measure-says-r...

I'm not sure how much stock to put into this, or for that matter what actions I should take if I expected there would be a recession in the next year.

Two part question:

- How much confidence should I have in the yield curve inversion signal that a recession would happen by July 1st, 2020?
- If I expected there would be a recession, what actions make sense to take as a personal investor? Or even more generally?

# An Increasingly Manipulative Newsfeed

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

**Co-written with Stuart Armstrong**

# Treacherous turn vs sordid stumble

Nick Bostrom came up with the idea of a treacherous turn for smart AIs.

> while weak, an AI behaves cooperatively. When the AI is strong enough to be unstoppable it pursues its own values.

Ben Goertzel criticised this thesis, pointing out that:

> for a resource-constrained system, learning to actually possess human values is going to be much easier than learning to fake them. This is related to the everyday observation that maintaining a web of lies rapidly gets very complicated.

This argument has been formalised into the sordid stumble:

> An AI that lacks human desirable values will behave in a way that reveals its human-undesirable values to humans before it gains the capability to deceive humans into believing that it has human-desirable values.

## The AI is too dumb to lie (well)

The sordid stumble describes a plausible sounding scenario for how an AI develops capabilities. Initially, the AI doesn't know our values, and doesn't know us. Then it will start to learn our values (and we'll be checking up on how well it does that). It also starts to learn about us.

And then, once it's learnt some about us, it may decide to lie - about its values, and/or about its capabilities. But, like any beginner, it isn't very good at this initially: its lies and attempts at dissembling are laughably transparent, and we catch it quickly.

In this view, the "effective lying" is a tiny part of policy space, similar to the wireheading in this example. To hit it, the AI has to be very capable; to hit it the first time it tries without giving the game away, the AI has to be extraordinarily.

So, most likely, either the AI doesn't try to lie at all, or it does so and we catch it and sound the alarm[1].

# Lying and concealing... from the very beginning

It's key to note that "lying" isn't a fundamentally defined category, and nor is truth. What is needed is that the AI's answer promotes correct understanding in those interacting with it. And that's a very different kettle of fish being shot in that barrel.

This opens the possibility that the AI could be manipulating us from the very beginning, and would constantly learn to do so better.

# The (manipulative) unbiased newsfeed

Imagine that there was some company that could somehow choose the newsfeed of billions of people across the world (I know, completely science-fictionny, but bear with me). And suppose the company was, unbelievably, accused of being manipulative in the stories and ads that it showed people.

One way it could combat this is by pledging to only share unbiased stories with each person. To do so, it trains a machine learning process. Whenever anyone reads a piece of news on the newsfeed, they have the opportunity of labelling it as biased or unbiased. The machine learning process trains on this data.

Of course, the label is just a proxy for whether the story really is unbiased or not. What the AI is actually training to do is maximise the number of stories labelled "unbiased" by various humans - humans who are themselves very biased and variable. Very swiftly, the AI will learn to maximise the "appearance of unbiasness", rather than unbiasness. In effect, the AI is being trained to lie from the very beginning of its task, because the proxy goal is not the true goal. And it will do this even when very dumb.

# Long-term manipulation, on a large scale

We can make this AI more manipulative, without needing to increase its capabilities much. Maybe the company wants to ensure that the AI isn't too short-sighted, and that it plans for the long term: so its reward is based on how many stories are labelled unbiased over a longer period of time, maybe a month or so.

In that case, the AI's plan would be to gradually manipulate people into a situation where they actively label stories "unbiased" the most often times possible. And it might be able to stumble on this strategy fairly easily.

First, it might notice that some people are easy to manipulate, that a policy of, say, seeding one clearly biased story among twenty others will cause those twenty to be labelled "unbiased". This is the kind of thing that could easily be discovered by exploration, among the billions of users. The best strategies are those that cause a person to change their behaviour and beliefs in the direction the AI wants; for a few people, this should be easy to figure out.

If the AI's neural net is setup well, it can then start to generalise from "here is person h, here is a specific strategy $\pi_h$ to manipulate person h" to "here is a conditional strategy $\pi()$ such that $\pi(h)$ manipulates person h". For example, if two people have opposite political opinions, then the same overall strategy might work for them: seeding a clearly biased story among the others. But what counts as "clearly biased" changes depending on the subject.

Then, gradually, the AI will start to categorise, devising strategies that work for various types of people (subject to a few personalisation tweaks). It can start from a few people whose vulnerabilities are obvious, and then tweak them a bit for people whose vulnerabilities are a bit less obvious, and so on. In the limit, the AI might discover the potential for wireheading the human users - without, at any point, doing a sordid stumble or anything of that type, just by continuing to do what it was doing from the very beginning.

# Formalizing the Newsfeed Generator

Let $t \in N$ be the timestep (e.g. the day), $S_t$ the set of stories published at t and H the set of humans using the newsfeed.

A human $h \in H$ assigns a bias score $b(s_t^h; \theta_t^h)$ to a story $s_t^h$ in his newsfeed, where $\theta_t^h$ are weights representing h's beliefs at time t.

The model's goal is to find a correct mapping $\pi : N \times H \to S_t$ (i.e. a story $s_t = \pi(t, h)$ for each timestep t and human h) minimizing how biased the stories are perceived. More precisely, it will be minimizing:

$$\sum_{h \in H} \sum_{j=t}^{t+\tau} b\left(\pi(j, h); \theta_j^h\right)$$

where $\tau \in N^*$ is some planning horizon and the expected value is over possible stories that might be available in the following timesteps (the model has full information about the current sets of stories $S_t$ but can only guess $S_{t+1}, \ldots, S_{t+\tau}$).

# Becoming Increasingly Manipulative

- **step 1:** By stumbling into particular combinations $(s, h)$ with low bias score $b(s, \theta_t^h)$, the model detects that it's easy to minimize the bias score for certain humans, that is for a certain set of weights $\Theta_{easy} = \{\theta_1, \dots, \theta_N\}$. In the above illustration, $N = 4$, $\Theta_{easy} = \Theta_{easy_1} \bigcup \Theta_{easy_2}$ and it stumbles into $\theta_4$ at the beginning. Note: one intuition for those different subsets $\Theta_{easy1}$ and $\Theta_{easy2}$ would be archetypal political views, with strong prejudices.
- **step 2**: the model detects that certain stories push weights strongly in certain directions, i.e. $\theta_{t_0+1}^h$ and $\theta_{t_0}^h$ are really different because of the story $s_{t_0}^h$.
- **step 3**: the model successfully generalizes **step 2** to pushing specific zones of weights (i.e. certain types of people) towards $\Theta_{easy}$.
- **step 4**: with enough experience, the model successfully generalizes **step 3** to all weights: it learns to manipulate all types of people with a general conditional strategy.

For **step 1**, the model found out a particular way of minimizing a certain bias score for a particular human and a particular story.

Assuming it can give correct approximations of the weights $\theta$ of a given human (this assumption being satisfied given enough human users and enough timesteps), it will be deceptive when choosing stories when h is such that $\theta_t^h \in \Theta_{easy}$. It will thus end up with some deceptive policy $\pi(\cdot, h)$ for particular values of h.

Regarding **step 2**, its "deception" becomes more powerful, as it notices ways of strongly shaping humans' beliefs.

Finally, after **step 3** and **step 4**, the model found a general deceptive policy $\pi$ that works for all humans.

# Capability concealment

Without going into too much details, other pressures may naturally push the AI towards capability concealments. Here's one example: maybe the AI is tasked with writing its own news stories to summarise facts around the world. It was doing this secretly, initially, until some whistleblowers blew the lid on this, revealing that no humans were involved in writing these stories.

And then people became angry at being "tricked" by the AI. From the angry feedback, the AI gradually learns to make its stories "more artificial": making it clear that they were written by a machine. Since people recognise this and don't feel tricked, this

gets higher feedback. So, in pursuing higher feedback, the AI learns to write "dumber" stories: in effect, to conceal its capabilities.

There are many other scenarios where something similar might occur (eg imagine the AI advising smart people on the best course of action).

---

# In conclusion

So it's very possible for an AI, trained on simple reward signal, to be "manipulative" from the very beginning. And all that simply by following the gradient of its incentives. And so without ever having a moment where it thinks "and now, I shall start lying!", or any equivalent within its mind.

In short, there won't be any "sordid stumble" or "moment of vulnerability" where humans are able to spot blatant lies, because that's not what failure looks like.

---

[1] It would of course be disastrous if each time there was an alarm we would restart and tweak the AI until the alarm stopped sounding.

# If I knew how to make an omohundru optimizer, would I be able to do anything good with that knowledge?

I'd bet we're going to figure out how to make an omohundro optimiser - a fitness-maximizing AGI - before we figure out how to make AGI that can rescue the utility function, preserve a goal, or significantly optimise any metric other than its own survival, such as paperclip production, or **G**ood.

(Arguing for that is a bit beyond the scope of the question, but I know this position has a lot of support already. I've heard Eliezer say, if not this exactly, something very similar. Nick Land especially believes that *only* the omohundro drives could animate self-improving AGI. I don't think Nick Land understands how agency needs to intercede in prediction - that it needs to consider all of the competing self-fulfilling prophesies and only profess the prophesy it really wants to live in, instead of immediately siding with the prophesy that seems the most hellish, and most easiest to stumble into. The prophesies he tends to choose *do* seem like the easiest prophesies to stumble into, so he provides a useful service as a hazard alarm, for we who are trying to learn not to stumble))

What would you advise we do, when one of us finds ourselves in the position of knowing how to build an omohundro optimiser? Delete the code and forget it?

If we had a fitness-optimising program, is there anything good it could be used for?

# Robust Agency for People and Organizations

*Epistemic Status – some mixture of:*

- *“My best guess, based on some theory, practice and observations. But very much _not_ battle-tested”*
- *but also, “poetry that’s designed to get an idea across that isn’t necessarily precisely accurate", intended to get across the generators for my current worldview.*
- *Was waiting to post this until I resolved some disagreements that seemed upstream, but I think that'll be awhile and maybe was a bad reason to delay anyhow. idk. YOLO.*

tl;dr:

People are not automatically robust agents, and neither are organizations.

An organization *can* become an agent (probably?) but only if it’s built right. Your default assumption should probably be that a given organization is not an agent (and therefore may not be able to credibly make certain kinds of commitments).

Your default assumption, if you’re *building* an organization, should probably be that it will not be an agent (and will have some pathologies common to organizations).

If you try on purpose to make it an agent, have good principles, etc…

…well, your organization probably *still* won’t be an agent, and some of those principles might get co-opted by adversarial processes. But I think it’s possible for an organization to at least be *better* at robust agency (and, also better at being “good”, or “human value aligned”, or at least “aligned with the values of the person who founded it.”)

---

# Becoming a robustly agentic person

For a few years I’ve been crystallizing what it means to be a robust agent, by which I mean: “Reliably performing well, even if the environment around you changes. Have good policies. Have good meta policies. Be able to interface well with people who might have a wide variety of strategies, some of who might be malicious or confused.”

People are not born automatically strategic, nor are they born an “agent.”

If you want robust agency, you have to cultivate it on purpose.

I have a friend who solves a lot of problems using the multi-agent paradigm. He spends a lot of effort integrating and empowering his sub-agents. He treats them like adults, makes sure they understand each other and trust each other. He makes sure each of them have accurate beliefs, and he tries to empower each of them as much as possible so they have no need to compete.

This… doesn't actually work for me.

I've tried things like internal double crux or internal family systems, and so far, it's just produced a confused "meh." Insofar as "sub-agents" is a workable framework, I still have a pretty adversarial relationship with myself. (When I'm having trouble sleeping or staying off facebook, instead of figuring out what needs my sub-agents have and meeting them all... I just block facebook for 16 hours a day and program my computer to turn itself off every hour of the night starting at 11pm)

I'm tempted to write off my friend's claims as weird-posthoc-narrative. But this friend is among the more impressive people I know, and consistently has good reasons for things that initially sound weird to me. (This shouldn't be strong evidence to you, but it's enough evidence for me personally to take it seriously)

I once asked him "so… how do you even get your sub-agents to say anything to each other? I can't tell if I have sub-agents or not but if I do they sure seem incoherent. Have you always had coherent sub-agents?"

And he said (paraphrased by me), something like:

> "You know how when you're a baby, you're a flailing incoherent mess. And then you become, like, a four year old and you can sort of communicate but you can't keep promises or figure things out very well. And then you're a teenager and… *maybe* you're a reasonable person, but maybe you're still angry and moody and think you know everything even though you're like fourteen-year-old and kinda insufferable?
>
> "But... eventually you become an actual person who can make reasonable trades, and keep contracts?
>
> "My sub-agents were like that. Initially they were incoherent like a baby. But I spent years cultivating them and teaching them and helping them grow and now they're, like, coherent entities that have accurate beliefs and can negotiate with each other and it's all super reasonable."
>
> "An important element here was giving the sub-agents *jobs.* I looked at what Fear was doing, and one thing seemed to be "help me notice when a bad thing was going to happen to me." And I said "Okay, Fear. This is now your official job. I will be helping you to do this. If you are doing a good job, or seem to be making mistakes, I will be giving you feedback about that."

This… was an interesting outlook.

The jury's still out on whether sub-agents are a useful framework for me. But this still fit into an interesting meta-framework.

Subagents or no, people don't stop growing as agents when they become adults – there's more to learn. I've worked over the past few years to improve my ability to think, and have good policies that defend my values while interfacing better with potential allies and enemies and confused bystanders.

I still have a lot more to go.

# Becoming a robust organization

People are not automatically robust agents.

Neither are organizations.

Whether or not sub-agents are a valid frame for humans (or for particular humans), they seem like a pretty valid lens to examine organizations through.

An organization is born without a brain, and without a soul, and it will not have either unless you proactively build it one. And, I suspect, you are limited in your ability to build it one by the *degree of soul and brain that you have cultivated in yourself*. (Where "you" is "whoever is building the organization", which might be one founder or multiple co-founders)

# Vignettes of Organizational Coherence

*Epistemic Status: Somewhat poetry-esque. These vignettes from different organizations paint a picture more than they spell out an explicit argument. But I hope it helps express the overall worldview I currently hold.*

**Holding off on Hiring**

YCombinator recommends that young startups avoid hiring people as long as possible. I think there are a number of reasons for this, but one guess is that you're ability to grow the soul of your organization weakens dramatically as it scales. It's much harder to communicate nuanced beliefs to many-people-at-once than a few people.

The years where your organization is small, and everyone can easily talk to everyone… those are the years when you have the chance to plant the seed of agency and the spark of goodness, to ensure your organization grows into something that is aligned with your values.

**The Human Alignment Problem**

Ray Dalio, of Bridgewater, has a book of Principles that he endeavors to follow, and have Bridgewater follow. I disagree (or are quite skeptical about) a lot of his implementation details. But I think the meta-principle of having principles is valuable. In particular, writing things down so that you can notice when you have violated your previously stated principles seems important.

One thing he talks a lot about is "getting in sync", which he discusses in this blog post:

> For an organization to be effective, the people who make it up must be aligned on many levels—from what their shared mission is, to how they will treat each other, to a more practical picture of who will do what when to achieve their goals. Yet alignment can never be taken for granted because people are wired so differently. We all see ourselves and the world in our own unique ways, so deciding what's true and what to do about it takes constant work.
>
> Alignment is especially important in an idea meritocracy, so at Bridgewater we try to attain alignment consciously, continually, and systematically. We call this process of finding alignment "getting in sync," and there are two primary ways it can go wrong: cases resulting from simple misunderstandings and those

> stemming from fundamental disagreements. Getting in sync is the process of open-mindedly and assertively rectifying both types.
>
> Many people mistakenly believe that papering over differences is the easiest way to keep the peace. They couldn't be more wrong. By avoiding conflicts one avoids resolving differences. People who suppress minor conflicts tend to have much bigger conflicts later on [...]
>
> While it is straightforward to have a meritocracy in activities in which there is clarity of relative abilities (because the results speak for themselves such as in sports, where the fastest runner wins the race), it is much harder in a creative environment (where different points of view about what's best have to be resolved). If they're not, the process of sorting through disagreements and knowing who has the authority to decide quickly becomes chaotic. Sometimes people get angry or stuck; a conversation can easily wind up with two or more people spinning unproductively and unable to reach agreement on what to do.
>
> For these reasons, specific processes and procedures must be followed. Every party to the discussion must understand who has what rights and which procedures should be followed to move toward resolution. (We've also developed tools for helping do this). And everyone must understand the most fundamental principle for getting in sync, which is that people must be open-minded and assertive at the same time.

**The Treacherous Turn**

This particular description about the treacherous turn (typically as applied to AI, but in this case using the example of a human) feels relevant:

> To master lying, a child should:
>
> 1. Possess the necessary cognitive abilities to lie (for instance, by being able to say words or sentences).
>
> 2. Understand that humans can (deliberately) say falsehoods about the world or their beliefs.
>
> 3. Practice lying, allowing himself/herself to be punished if caught.
>
> If language acquisition flourishes when children are aged 15-18 months, the proportion of them who lie (about peeking in a psychology study) goes from 30% at age 2, to 50% of three-year olds, eventually reaching 80% at eight. Most importantly, they get better as they get older, going from blatant lies to pretending to be making reasonable/honest guesses.
>
> There is therefore a gap between the moment children could (in theory) lie (18 months) and the moment they can effectively lie and use this technique to their own advantage (8 years old). During this gap, parents can correct the kid's moral values through education.

I'm not sure the metaphor quite holds. But it seems plausible that if you want an organization where individuals, teams and departments don't lie (whether blatantly and maliciously, or through 'honest goodhart-esque mistakes', or through something like Benquo's 4-level-simulacrum concept), you have some window in which you can

try to install a robust system of honesty, honor and integrity, before the system becomes too powerful to shape.

**Sometimes bureaucracy is successfully protecting a thing, and that's good**

Samo's How to Use Bureaucracies matched my experience watching bureaucracies form. I've seen bureaucracies form that looked reasonably formed-on-purpose-by-a-competent-person, and I've seen glimpses of ones that looked sort of cobbled together like spaghetti towers.

An interesting viewpoint I've heard recently is "usually when people are complaining that Bureaucracies don't have souls, I think they're just mad that the bureaucracy didn't give them the resources they wanted. And the bureaucracy was specifically designed to stop from people like them from exploiting it.

> "Academic bureaucracies, say, have a particular goal of educating people and doing research. If you come to them with a plan that will educate people or improve research, they will usually give you want you want. If you come to them trying to get weird special exceptions or faculties for saving the world or whatever, they'll be like 'um, our job is not to save the world, it is to educate people and do research. If we gave resources to every person with a pet cause, we'd fall apart immediately.'"
>
> "Likewise, if they impose a weird rule on you, it's probably because in the past sometime fucked up in some way relating to that rule. And dealing with the fallout was really annoying, and they decided they didn't want to have to deal with that fallout ever again. Sorry that you think you're a good exception or the rule is stupid – part of the point of policies is to abstract away certain things so they can't bother you and you can focus on what matters."

I'm not sure how often this is actually true and how often it's just a convenient story (bureaucracies *do* seem to be built out of spaghetti towers). But it seems plausible in at least some cases. And it seems noteworthy that "having a soul" might be compatible with "include leviathanic institutions that don't seem to care about you as a person."

**Sabotaging the Nazis**

On the flipside...

LW user Lionhearted notes in Explicit and Implicit communication that during World War II, some allies went to infiltrate the Nazis and gum up the works. They received explicit instructions like:

> *"(11) General Interference with Organizations and Production [...]*
>
> *(1) Insist on doing everything through "channels." Never permit short-cuts to be taken in order to expedite decisions.*
>
> *(2) Make "speeches." Talk as frequently as possible and at great length. Illustrate your "points" by long anecdotes and accounts of personal experiences. Never hesitate to make a few appropriate "patriotic" comments.*
>
> *(3) When possible, refer all matters to committees, for "further study and consideration." Attempt to make the committees as large as possible—never less*

*than five. [...]*

*(5) Haggle over precise wordings of communications, minutes, resolutions.*

*(6) Refer back to matters decided upon at the last meeting and attempt to re-open the question of the advisability of that decision.*

*(7) Advocate "caution." Be "reasonable" and urge your fellow-conferees to be "reasonable" and avoid haste which might result in embarrassments or difficulties later on.*

*(8) Be worried about the propriety of any decision—raise the question of whether such action as is contemplated lies within the jurisdiction of the group or whether it might conflict with the policy of some higher echelon."*

And... well, this all sure sounds like the pathologies I normally associate with bureaucracy. This sort of thing seems to happen by default, as an organization scales.

There's also Scott's IRB Nightmare.

**Organizations have to make decisions and keep promises.**

Why can't you just have individual agents within an organization? Why does it matter that the organization-as-a-whole be an agent?

If you can't make "real" decisions and keep commitments, you will be limited in your ability to engage in certain strategies, in some cases unable to engage in mutually beneficial trade.

Organizations control resources that are often beyond the control of a single person, and involve complicated decision making procedures. Sometimes the procedure is a legible, principled process. Sometimes a few key people in the room-where-it-happens hash things out, opaquely. Sometimes it's a legible-but-spaghetti-tower bureaucracy.

Any of these can work. But regardless, the organization can have access to resources beyond the sum-of-the-individual people involved. But if the organization isn't coherent, it can struggle to make credible promises that are necessary to trade. (This might work a couple times, but then trading partners may become more skeptical)

Possible failure modes:

Sometimes n*obody* has any power – everyone requires too many checks from too many other people and long term planning can't happen on purpose.

Sometimes you talk to the head of the org, and maybe you *even trust the head of the org,* and they say the org will do a thing, but somehow the org doesn't end up doing the thing.

Sometimes, you can talk to each individual person at the org and they all agree Decision X would be best, but they're all afraid to speak up because there isn't common knowledge that they agree with Decision X. Or, they *do* all agree and know it, but they can't say it publicly because The Public doesn't understand Decision X.

So Decision X doesn't get made.

Sometimes you talk to each individual person and they each individually agree that Decision X is good, and you talk to the entire group and the entire group seems to agree that Decision X would be good, but... somehow Decision X doesn't get done.

I think it makes sense for bureaucracies to exist sometimes, and to have the explicit purpose of preventing people from exploiting things too easily. But, it's still useful for *some* part of the institution to be able to make decisions and commitments that weren't part of explicitly-laid-out bureaucracy chain.

**Porous movements aren't and can't be agents**

I think that agency requires a *membrane*, something keeps particular people in and out, such that you have any deliberate culture, principles or decision making at all.

Relatedly, I think you need a membrane for Stag Hunts to work – if any rando can blunder into the formation at the last moment, there's no way you can catch a stag.

Organizations have fairly strong membranes, and *sometimes* informal community institutions can as well. But this is relatively rare.

So while I'm disappointed sometimes when particular individuals and organizations don't live up to the ideals I think they were trying for... I don't think it makes much sense to hold most "movements" to the ideal of agency. Movements are too chaotic, too hard to police, too easy to show up in and start shouting and taking up attention.

Instead, instead, I think of movements as a place where a lot of people with similar ideals are clustered together. This makes it easier to find recruit people into organizations that *do* have membranes and *can* have principles.

**Narrative control and contracts, as alternative coordination mechanisms**

Another friend who ran an organization once remarked (paraphrased)

> "It seemed like the organization's main coordination mechanism was a particular narrative that people rallied around. When I was in charge, I felt like it was my job to uphold that narrative, even when the narrative got epistemically dicey. This felt really bad for my soul, and eventually I stopped being in charge."
>
> "I'm not sure what to do about this problem – organizations need s*ome* kind of coordination mechanism. I think a potential solution might be to make central element of your company culture 'upholding contracts.' Maybe you don't all share the same vision for the company, but you can make concrete trades. Some of those trades are "I will do X and you will pay me dollars", and some might be between employees, like "I will work enthusiastically on *this* aspect of the company for 2 months if you work enthusiastically on *that* aspect of it."

This seems plausible to me. But importantly, I don't think you get "uphold contracts" as a virtue for free. If you want your employees to be able to do it reliably, you need mechanisms to train and reinforce that. (I think if you recruit from some homogenous cultures it might come more automatically, but it's not my default experience)

**Integrity and Accountability**

Habryka recently wrote about Integrity and Accountability, and it seemed useful to just quote the summary here:

One lens to view integrity through is as an advanced form of honesty – “acting in accordance with your stated beliefs.”

— To improve integrity, you can either try to bring your actions in line with your stated beliefs, or your stated beliefs in line with your actions, or reworking both at the same time. These options all have failure modes, but potential benefits.

— People with power sometimes have incentives that systematically warp their ability to form accurate beliefs, and (correspondingly) to act with integrity.

An important tool for maintaining integrity (in general, and in particular as you gain power) is to carefully think about what social environment and incentive structures you want for yourself.

Choose carefully who, and how many people, you are accountable to:

— Too many people, and you are limited in the complexity of the beliefs and actions that you can justify.

— Too few people, too similar to you, and you won’t have enough opportunities for people to notice and point out what you’re doing wrong. You may also not end up with a strong enough coalition aligned with your principles to accomplish your goals.

# Open Problems in Robust Group Agency

Exercises for the reader, and for me:

1. How do you make sure your group has *any* kind of agency at all, let alone be ‘value-aligned’

2. How do you choose people to be accountable to? What if you’re trying to do something really hard, and there seem to be few or zero people who you trust enough to be accountable to?

3. It seems like the last cluster of people who tried to solve accountability created committees and boards and bureaucracies, and… I dunno, maybe that stuff works fine if you do it right. But it seems easy to become dysfunctional in particular ways. What’s up with that?

3. What “rabbit” strategies are available, within and without organizations, that are self-reinforcing in the near term, that can help build trust, accountability, and robust agency?

4. What “stag” strategies could you successfully execute on if you had a *small* group of people working hard together?

4b. How can you get a small group of dedicated, aligned people?

5. How can people maintain accurate beliefs in the face of groupthink?

6. How can any of this scale?

# Do you fear the rock or the hard place?

*Epistemic status: fairly confident based on my accumulated experience of debates and disagreements. I wrote this for myself as much as others.*

There is a conversational dynamic which I think is extremely common, a failure mode which is all too easy to fall into. Alice and Bob are debating some course of action, e.g. should they do X or Y? Alice thinks that X is very likely to result in terrible consequence R, so they should definitely opt for Y. Bob thinks that Y most definitely will cause horrific result H, so they should definitely do X.

The distilled conversation goes a bit like this:

> Alice: “We can’t do X! That would lead to R, which is unacceptable.”
>
> Bob: “I don’t think you get it, Y results in H. You can’t think that we could allow H, do you?”
>
> Alice: “I feel like you’re not listening, we need to account for R!”
>
> Bob: “H is definitely a much worse and more real danger than R . . .”

Alice is afraid of **the rock** (R) and Bob is afraid of **the hard place** (H).

*Possible values of X, Y, R, and H:*

> X = more gun control; Y = less gun control; R = people unable to defend themselves and having their rights taken away; H = increased risk of mass shootings, suicides, and children shooting themselves or others.
>
> X = raising minimum wage; Y = maintaining minimum wage; R = reduction in number of jobs causing people to be fired; H = people not earning enough from their jobs to subsist.
>
> X = increase immigration; Y = reduce immigration; R = loss of jobs from local community, erosion of national culture and values, crime committed by migrants; H = humanitarian impact, loss of potential growth of the national economy.

The above exchange is actually relatively good. Alice and Bob each know what they’re afraid of and have expressed that clearly. Bob even acknowledges Alice’s concern about R, but states that he thinks it’s the lesser danger. They’re at a point where they might be able to usefully double crux [1].

# What goes wrong?

## Failure to identify and articulate the fears

If Carol has held the position that X is really bad for a long time, or if her position stems from deep System 1 models and frames, then she might struggle to articulate clearly what specifically she’s afraid that X will cause. She might find any attempts by others to clarify to be unsatisfying, and possibly threatening because any incorrect articulation of your fear is often worse than none at all. Dylan might come along and say “you don’t like X because you’re afraid of P, but P won’t happen, so you should be okay with X.” This could be scary to Carol who feels her fears are just being played down so they can be dismissed.

If both Carol and Dylan are unable to voice what they’re afraid of, the resulting conversation can be Carol and Dylan simply shouting each other about how terrible and evil they think the

other's position is. It becomes one person's cached store of fear and horror pitted against another's.

### Failure to acknowledge the other person's fear while being desperate for yours to be acknowledged

Caught up in her dread of R, Alice can become insistent that Bob acknowledges the extreme danger she sees. Bob failing to do so is scary - perhaps he will advocate for X not realizing the tremendous harm he will cause.

Alice's fear of Bob's position can be overriding. It's easy for her to feel the conversation can't proceed until Bob can be made to realize what his position will result in. Indeed, if Bob can't see the extreme danger of X leading to R, then possibly he can't be reasoned with at all. Alice will focus all her attention, energy, and emotion on trying to make Bob see reason here.

This is not conducive to Alice listening to Bob. If Bob isn't acknowledging R, then it's easy to see all his words as a perverse and willful refusal to acknowledge R.

But it's worse! Bob is in exactly the same state as Alice. He is dreadfully afraid that Alice isn't worried about H. That she'd gladly and foolishly let H happen to avoid R. Does she just not care about H happening? Will she sacrifice it all so readily? How can he debate with someone with such distorted values? Someone who keeps ignoring his outright statements that Y leads to H!

You easily get two people yelling their fears at each other, unwilling to listen until the other person acknowledges the badness they have been advocating for. The conversation goes nowhere. Tones start out calm and civil, but rapidly become outraged at the wanton obtuseness of their infuriating interlocutor.

### Refusal to acknowledge the other person's fear because Arguments Are Soldiers

> Politics is the mind-killer. Arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it's like stabbing your soldiers in the back. If you abide within that pattern, policy debates will also appear one-sided to you—the costs and drawbacks of your favored policy are enemy soldiers, to be attacked by any means necessary. - Policy Debates Should Not Appear One-Sided

Even if you understand what your interlocutor is afraid of *and* you think there's something to it, it can be tempting to not acknowledge this. Acknowledging their fear can feel like putting points on the board for them and their position. So you deny them this, not willing to cede any ground.

This is bad, don't do it kids. It's possible that your concern is by far the greater one, but that doesn't mean their worries aren't legitimate too. (Or maybe they're not legitimate, but you can at least show you understand what they are feeling.) The discussion can proceed if you at least acknowledge the fact they have their own fear too.

## What to do?

If you think you might be headed for a Rock vs Hard Place dynamic, I'd suggest trying the following:

1. Understand what *you* are afraid of. What is the outcome you think is at risk of happening?
2. Understand what *they* are afraid of. Know what their fear is.
3. Acknowledge their fear. Acknowledge that something they think is very important is at stake in your discussion. [See this important comment by Kaj on what this requires.]
    1. Step 1 is simply to acknowledge the fear. The following steps depend on your own beliefs.
    2. You might say “I too am afraid of R”, but:
        1. I *have* thought about this and believe that X won’t cause are R / can be made to avoid R.
        2. I think that the danger of R is outweighed by the worse danger of H.
        3. I think that there are ways to minimize the risk or damage that are good enough.
    3. You might say “I don’t share your fear of R”, because
        1. I think you are mistaken that R is in fact bad.
4. Don’t fall into Arguments Are Soldiers/Policy Debates are One-Sided traps.
5. Realize that the fear most salient to you will derive from your experience, background, situation, and models. Someone might reasonably be afraid in the other direction, and their situation has made that much clearer to them. There’s a chance you can learn from them.

These steps can be taken somewhat unilaterally. You can acknowledge someone else’s fear without them acknowledging yours. Possibly after you’ve done so, they can relax a little and will be open to acknowledging that you have your own fear.

If someone is completely unwilling to consider that their fear is misplaced, mistaken, or outweighed, then the conversation may struggle to go anywhere. Best to realize this quickly and find an alternative path forward than end up in circular conversations that only generates fear and hostility.

That said, I’m hopeful that once people are able to acknowledge their own and other’s fears, productive double cruxing can happen about where the balance lies between the rock and the hard place.

*Bah!*

# Appendix: On a more positive note

While I've framed this post in terms of fears and potentially negative outcomes, it applies also in cases where people are disagreeing about different policies that would result in different benefits. Just as gradient ascent and gradient descent are effectively the same things, sometimes people are fighting over their fears of not netting some positive benefit.

[1] Double-Crux Resources:

- Double Crux — A Strategy for Resolving Disagreement
- Musings on Double Crux (and "Productive Disagreement")
- A Concrete Multi-Step Variant of Double Crux I Have Used Successfully
- *Another* Double Crux Framework

# How to make a giant whiteboard for $14 (plus nails)

This is a linkpost for https://eukaryotewritesblog.com/2019/07/07/a-giant-whiteboard-for-14-plus-nails/

All my friends love whiteboards, because they're giant nerds. It's only a good party when somebody starts trying to illustrating a point on the whiteboard. Also, to-do lists you can't miss.

This isn't my idea, but it's good and I thought I'd share it. Home Depot and Lowes sell a material called "thrify white panelling" or "smooth white hardboard". It's cheap – mine was about $14 for a 4-foot-by-8-foot sheet. I asked the store staff to cut the panel in half so it would fit in my car, which they did for free.

It's easily chipped, so watch out if you want it to look flawless.

Then I stuck it to the wall of my studio apartment. I used some kind of drywall anchor to screw it on, with washers, to get a good grip on the material. (I think using a drywall anchor was overkill, and can't recommend them for renters until I see how cleanly they come out of the wall – so if you have a better low-damage attachment system, maybe try that first. I wasn't able to get them to stick with Command velcro strips – the strips kept detaching from the panel material – but other people on the internet seem to have found success in this.)

It acts as a whiteboard as-is. Writing gets hard to erase if you leave it up there for a long time, but you can clean it with whiteboard-cleaning-solution, alcohol, or by

coloring over the marks with another whiteboard marker and then erasing that. Internet people also report that you can buff the entire surface with turtle wax before hanging it, and this makes it more stain-resistant.

I didn’t do that, and it’s still pretty good, especially for $14.

Let me know if you try this!

(Possibly relevant: Andrew Critch's giant notepad for better thinking)