Best of LessWrong: October 2017

1. Robustness as a Path to AI Alignment
2. Inadequacy and Modesty
3. Slack for your belief system
4. There's No Fire Alarm for Artificial General Intelligence
5. Four Scopes Of Advice
6. Against naming things, and so on
7. The Problematic Third Person Perspective
8. I Can Tolerate Anything Except Factual Inaccuracies
9. Writing That Provokes Comments
10. Multidimensional signaling
11. Windows Resource Repository
12. That's a Thing!
13. AlphaGo Zero and the Foom Debate
14. Community Capital
15. Tensions in Truthseeking
16. "Focusing," for skeptics.
17. Social Choice Ethics in Artificial Intelligence (paper challenging CEV-like approaches to choosing an AI's values)
18. Placing Yourself as an Instance of a Class
19. Trope Dodging
20. De-Centering Bias
21. The Strengths of the Two Systems of Cognition
22. Different Worlds
23. Tech vs. Willpower
24. Infant Mortality and the Argument from Life History
25. HOWTO: Screw Up The LessWrong Survey and Bring Great Shame To Your Family
26. Seek Fair Expectations of Others' Models
27. Time to Exit the Sandbox
28. Geeks, Mops, Sociopaths
29. What would convince you you'd won the lottery?
30. Distinctions in Types of Thought
31. Leaders of Men
32. Yudkowsky on AGI ethics
33. Norms For Link Posts
34. What Evidence Is AlphaGo Zero Re AGI Complexity?
35. Prosocial manipulation
36. Winning is for Losers
37. Frequently Asked Questions for Central Banks Undershooting Their Inflation Target
38. Outline of an approach to AGI Estimation
39. Fighting the evil influence of Facebook (but keeping the good bits): a manifesto and how-to guide
40. In defence of epistemic modesty
41. Things you should never do
42. Contra double crux
43. Why no total winner?
44. The Typical Sex Life Fallacy
45. Beginners' Meditation
46. Bring Back the Sabbath
47. Unofficial ESPR Post-mortem
48. React and Respond
49. Avoiding Selection Bias
50. Identities are [Subconscious] Strategies

Best of LessWrong: October 2017

1. Robustness as a Path to AI Alignment
2. Inadequacy and Modesty
3. Slack for your belief system
4. There's No Fire Alarm for Artificial General Intelligence
5. Four Scopes Of Advice
6. Against naming things, and so on
7. The Problematic Third Person Perspective
8. I Can Tolerate Anything Except Factual Inaccuracies
9. Writing That Provokes Comments
10. Multidimensional signaling
11. Windows Resource Repository
12. That's a Thing!
13. AlphaGo Zero and the Foom Debate
14. Community Capital
15. Tensions in Truthseeking
16. "Focusing," for skeptics.
17. Social Choice Ethics in Artificial Intelligence (paper challenging CEV-like approaches to choosing an AI's values)
18. Placing Yourself as an Instance of a Class
19. Trope Dodging
20. De-Centering Bias
21. The Strengths of the Two Systems of Cognition
22. Different Worlds
23. Tech vs. Willpower
24. Infant Mortality and the Argument from Life History
25. HOWTO: Screw Up The LessWrong Survey and Bring Great Shame To Your Family
26. Seek Fair Expectations of Others' Models
27. Time to Exit the Sandbox
28. Geeks, Mops, Sociopaths
29. What would convince you you'd won the lottery?
30. Distinctions in Types of Thought
31. Leaders of Men
32. Yudkowsky on AGI ethics
33. Norms For Link Posts
34. What Evidence Is AlphaGo Zero Re AGI Complexity?
35. Prosocial manipulation
36. Winning is for Losers
37. Frequently Asked Questions for Central Banks Undershooting Their Inflation Target
38. Outline of an approach to AGI Estimation
39. Fighting the evil influence of Facebook (but keeping the good bits): a manifesto and how-to guide
40. In defence of epistemic modesty
41. Things you should never do
42. Contra double crux
43. Why no total winner?
44. The Typical Sex Life Fallacy
45. Beginners' Meditation
46. Bring Back the Sabbath

47. Unofficial ESPR Post-mortem
48. React and Respond
49. Avoiding Selection Bias
50. Identities are [Subconscious] Strategies

Robustness as a Path to AI Alignment

[Epistemic Status: Some of what I'm going to say here is true technical results. I'll use them to gesture in a research direction which I think may be useful; but, I could easily be wrong. This does not represent the current agenda of MIRI overall, or even my whole research agenda.]

Converting Philosophy to Machine Learning

A large part of the work at MIRI is to turn fuzzy philosophy problems into hard math. This sometimes makes it difficult to communicate what work needs done, for example to math-savvy people who want to help. When most of the difficulty is in finding a problem statement, it's not easy to outsource the intellectual labor.

Philosophy is also hard to get traction on. Arguably, something really good happened to the epistemic norms of AI research when things switched over from GOFAI to primarily being about machine learning. Before, what constituted progress in AI was largely up to personal taste. After, progress could be verified by achieving high performance on benchmarks. There are problems with the second mode as well -- you get a kind of bake-off mentality which focuses on tricks to get higher performance, not always yielding insight. (For example, top-performing techniques in machine learning competitions often combine many methods, taking advantage of the complementary strengths and weaknesses of each. However, this approach leans on the power of the other methods.) Nonetheless, this is better for AI progress than armchair philosophy and toy problems.

It would be nice if AI alignment could be more empirically grounded. There are serious obstacles to this. Many alignment concerns, such as self-modification, barely show up or seem quite easy to solve when you aren't dealing with a superintelligent system. However, I'll argue that there is sometimes a way to turn difficult alignment problems into machine learning problems.

A second reason to look in this direction is that in order to do any good, alignment research has to be used by the people who end up making AGI. The way things look right now, that means they have to be used by machine learning researchers. To that end, anything which puts things closer to a shape which ML researchers are familiar with seems good.

To put it a different way: in the end, we need a successful pipeline from philosophy, to math, to implementation. So far, MIRI has focused on optimizing the first part of that pipeline. I think it may be possible to do research in a way which helps optimize the second part.

We wouldn't want the research direction to be constrained by this, since in the end we need to figure out what actually works, not what creates the most consumable research. However, I'll also argue that the research direction is plausible in itself.

The Big Picture

I started working full-time at MIRI about three months ago. In my second week, we had a research retreat in which we spent a lot of time re-thinking how all of the big research problems connect with each other and to the overall goal. I came out of this with the view that things factored somewhat cleanly into three research areas: *value learning*, *robust optimization*, and *naturalized agency*. [Again, this write-up isn't intended reflect the view of MIRI as a whole.]

1. **Value Learning:** The first problem is to specify what is "good" or what you "want" in enough detail that nothing goes wrong when we optimize for it. This is too hard (since humans seem really bad at knowing what they want in precise terms), so it would be nice to reduce it to a learning problem, if possible. This requires things like learning human concepts (including the concept "human") and accounting for bounded rationality in learning human values (so that you don't assume the human *wanted* to stub its toe on the coffee table).

2. **Robust Optimization:** We will probably get #1 wrong, so how can we specify systems which don't go off-track too badly if their values are misspecified? This includes things like transparency, corrigibility, and planning under moral uncertainty (doing something other than max-expected-value to avoid over-optimizing). Ideally, you want to be able to ask a superintelligent AI to make burritos, and *not* end up with a universe tiled with burritos. This corresponds approximately to the AAMLS agenda.

3. **Naturalized Agency:** Even if we just *knew* the correct value function and knew how to optimize it in a robust way, we don't actually know how to build intelligent agents which optimize values. It's a bit like the difference between knowing that you want to classify images and getting to the point where you optimize neural nets to do so: you have to figure out that squared-error loss plus a regularizer works well, or whatever. We aren't to the point where we just know what function to optimize neural nets for to get AGI out, value-aligned or no. Existing decision theories, agent frameworks, and definitions of intelligence don't seem up to the task of examining what rational agency looks like when the agent is embedded in a world which is bigger than it (so the real world is certainly not in the hypothesis space which the agent can represent), the agent can self-modify (so reflective stability and self-trust becomes important), and the agent is part of the world (so agents must understand themselves as physics and consider their own death).

To storify: *AI should do X such that X=argmax(value(x))!* ***WAIT!*** *We don't know what value is! We should figure that out!* ***WAIT! T****rying to argmax a slightly wrong thing often leads to more-than-just-slightly wrong results! We should figure out some other operation than argmax, which doesn't have that problem! WAIT! The universe isn't actually in a functional form such that we can just optimize it! What are we supposed to do?*

In a sense, this is a series of proxy problems. We actually want #1, but we've done relatively little on that front, because it seems much too confusing to make progress on. #2 still cuts relatively close to the problem, and plausibly, solving #2 means not needing to solve #1 as well. More has been done on #2, but it is still harder and more confusing than #3. #3 is fairly far removed from what we want, but working on #3 seems to plausibly be the fastest route to resolving confusions which block progress on #1 and #2.

What I want to outline is a particular way of thinking which seems to be associated with progress on both #2 and #3, and which also seems like a good sign for the philosophy→implementation pipeline.

What is Robustness?

(I'm calling this thing "robustness" in *association* with #2, but "robust optimization" should be thought of as its own thing -- robustness is necessary for robust optimization, but perhaps not sufficient.)

Robustness might be intuitively described as tolerance to errors. Put in a mathematical context, we can model this via an adversary who has some power to trip you up. A robustness property says something about how well you do against such adversaries.

For example, take quantilization. We want an alternative to optimization which is robust to misspecified utility functions. A Bayesian approach might introduce a probability distribution over possible utility functions, and maximize expected utility with respect to that uncertainty. This doesn't do much to increase our confidence in the outcome; we've only pushed the problem back to correctly specifying our uncertainty over the utility distribution, and problems from over-optimizing a misspecified function seem just about as likely. Certainly we get no new formal guarantees.

So, instead, we model the situation by supposing that an adversary has some bounded amount of power to deceive you about what your true utility function is. The adversary might concentrate all of this on one point which you'll be very mistaken about (perhaps making a very bad idea look very good), or spread it out across a number of possibilities (making many good ideas look a little worse), or something inbetween. Under this assumption, a policy which randomizes actions somewhat rather than taking the max-expected-utility action is effective. This gives you some solid guarantees against utility misspecification, unlike the naive Bayesian approach. There's still more to be desired, but this is a clear improvement.

Mathematically, an adversarial assumption is just a "for all" requirement. Bayesians are more familiar with doing well *in expectation*. Doing well in expectation has its merits. However, adversarial assumptions create stronger guarantees on performance, by optimizing for the worst case.

Garrabrant Inductors as Robustness

Garrabrant Inductors (AKA logical inductors) are MIRI's big success in naturalized agency. (Reflective oracles come in second.) They go a long way to clear up confusions about logical uncertainty, which was one of the major barriers to understanding naturalized agents. When I say that there has been more progress on naturalized agency than on robustness, they're a big part of the reason. Yet, at their heart is something which looks a lot like a robustness result: the logical induction criterion. You take the set of all poly-time trading strategies on a belief market, and ask that a Garrabrant inductor doesn't keep losing against any of these forever. This is very typical of bounded-loss conditions in machine learning.

In return, we get reliability guarantees. Sub-sequence optimality means that we get the benefits of the logical induction criterion no matter which subset of facts we actually care about. Calibration means that the beliefs can be treated as frequencies, and unbiasedness means these frequencies will be good even if the proof system is biased (selectively showing evidence on one side more often than the other). Timely learning means (among other things) that it doesn't matter too much if the theorem prover we're learning from is slow; we learn to predict things as quickly as possible (eventually).

The logical induction criterion is a relaxation of the standard Bayesian requirement that there be no Dutch Book against the agent. So, the Dutch Book argument for the axioms of probability theory has an adversarial form as well. The same can be said of the money-pump argument which justifies expected utility theory. Bayesians are not so averse to adversarial assumptions as they may seem; lurking behind the very notion of "doing well in expectation" is a "for all" requirement! Bayesians turn up their noses at decision procedures which try to do well in any other than an average-case sense because they *know* such a procedure is money-pumpable; an adversary could swoop in and take advantage of it!

This funny mix of average-case and worst-case reasoning is at the very foundation of the Bayesian edifice. I'm still not quite sure what to think of it, myself. Philosophically, what should determine when I prefer an average-case argument vs a worst-case one? But, that is a puzzle for another time. The point I want to make here is that there's a close connection between the types of arguments we see at the foundations of decision theory (Dutch Book and money-pump arguments which justify notions of rationality in terms of guarding yourself against an adversary) and arguments typical of machine learning (bounded-loss properties).

The Dutch Book argument forces a tight, coherent probability distribution, which can't be both computable and consistent with logic. Relaxing things a little yields a wealth of benefits. What other foundational arguments in decision theory can we relax a little to get rich robustness results?

Path-Independence

These examples are somewhat hand-wavy; what I'll say here is true, but hasn't yet brought forth any fruit in terms of AI alignment results. I am putting it here merely to provide more examples of being able to frame decision-theory things as robustness properties.

I've mentioned the Dutch Book argument. Another of the great foundational arguments for Bayesian subjective probability theory is Cox's Theorem. One of the core assumptions is that if a probability can be derived in many ways, the results must be equal. This is related to (but not identical with) the fact that it doesn't matter what order you observe evidence in; the same evidence gives the same conclusion, every time.

Putting this into an adversarial framework, this means the class of arguments which we accept doesn't leave us open to manipulation. Garrabrant Induction weakens this (conclusions are not fully independent of the order in which evidence is presented), but also gets versions of the result which a Bayesian can't, as mentioned in the previous section: it arrives at unbiased probabilities even if it is shown a biased

sampling of the evidence, so long as it keeps seeing more and more. (This is part of what saves Garrabrant Induction from my All Mathematicians are Trollable result.)

Another example illustrating the need for path-independence is Pascal's Mugging. If your utility function is unbounded and your probability distribution is something like the Solomonoff distribution, it's awfully hard to avoid having divergent expected utilities. What this means is that when you try to sum up your expected utility, the sum you get is dependent on the order you sum things in. This means Pascal can alter your end conclusion by directing your attention to certain possibilities, extracting money from you as a result.

It seems to me that path-independent reasoning is a powerful rationality tool which I don't yet fully understand.

Nuke Goodhart's Law From Orbit

(Repeatedly. It won't stay down.)

Goodhart's Curse is not the *only* problem in the robust optimization cluster, but it's close; the majority of the problems there can be seen as one form or another of Goodhart. Quantilizers are significant progress against goodhart, but not total.

1. Quantilizers give you a knob you can turn to optimize softer or harder, without clear guidance on how much optimization is safe. If you keep turning up the knob and seeing better results, what would make you back off from cranking it up as far as you can go?

2. Along similar lines, but from the AIs perspective, there's nothing stopping a quantilizer from building a maximizer in order to solve its problem. In our current environment, "implement a superintelligent AI to solve the problem" is far from the laziest solution; but in an environment containing highly intelligent quantilizers, the tools to do so are lying around. It can do so merely by "turning up its own knob".

Nonetheless, it seems plausible that progress can be made via more robustness results in a similar direction.

Something which has been discussed at MIRI, due to Paul Christiano's thoughts on the subject, is the Benign Induction problem. Suppose that you have some adversarial hypotheses in your belief mixture, which pose as serious hypotheses and make good predictions much of the time, but are actually out to get you; after amassing enough credibility, at a critical juncture they make bad predictions which do you harm.

One way of addressing this, inspired by the KWIK learning framework, is the consensus algorithm. How it works is, you don't output any probability at all unless your top hypotheses *agree* on the prediction; not just on the classification, but on the probability to within some acceptable epsilon tolerance. This acts as an honesty amplifier. Suppose you have a hundred hypotheses, and only one is good; the rest are corrupt. Even if the corrupt hypotheses can coordinate with each other, the one good hypothesis keeps them all in check: nothing they say gets out to the world unless they agree very closely with the good one. (However, they *can* silence the good hypothesis selectively, which seems concerning!)

A solution to the benign induction problem would be significant progress on the robust optimization problem: if we could trust the output of induction, we could use it to predict what optimization techniques are safe! (There's much more to be said on this, but this is not the article for it.)

So, quantilizers are robust to adversarial noise in the utility function, and the consensus algorithm is (partially) robust to adversarial hypotheses in its search space. Imagine a world where we've got ten more things like that. This seems like significant progress. Maybe then we come up with a Robust Optimization Criterion which implies all the things we want!

Machine learning experts and practitioners alike are familiar with the problems of over-optimization, and the need for regularization, in the guise of overfitting. Goodhart's Curse is, in a sense, just a generalization of that. So, this kind of alignment progress might be absorbed into machine learning practice with relative ease.

Limits of the Approach

One problem with this approach is that it doesn't provide *that* much guidance. My notion of robustness here is extremely broad. "You can frame it in terms of adversarial assumptions" is, as I noted, equivalent to "use *for all*". Setting out to use universal quantifiers in a theory is hardly much to go on!

It's not nothing, though; as I said, it challenges the Bayesian tendency to use "in expectation" everywhere. And, I think adding adversarial assumptions is a good brainstorming exercise. If a bunch of people sit down and try to come up with new parts to inject adversarial assumptions into for five minutes, I'm happy. It just may be that someone comes up with a great new robustness idea as a consequence.

Inadequacy and Modesty

The following is the beginning of *Inadequate Equilibria*, a new sequence/book on a generalization of the notion of efficient markets, and on this notion's implications for practical decision-making and epistemic rationality.

This is a book about two incompatible views on the age-old question: "When should I think that I may be able to do something *unusually well*?"

These two viewpoints tend to give wildly different, nearly *cognitively nonoverlapping* analyses of questions like:

- My doctor says I need to eat less and exercise, but a lot of educated-sounding economics bloggers are talking about this thing called the "Shangri-La Diet." They're saying that in order to lose weight, all you need to do is consume large quantities of flavorless, high-calorie foods at particular times of day; and they claim some amazing results with this diet. *Could they really know better than my doctor? Would I be able to tell if they did?*

- My day job is in artificial intelligence and decision theory. And I recall the dark days before 2015, when there was plenty of effort and attention going into advancing the state of the art in AI capabilities, but almost none going into AI alignment: better understanding AI designs and goals that can safely scale with capabilities. Though interest in the alignment problem has since increased quite a bit, it still makes sense to ask whether *at the time* I should have inferred from the lack of academic activity that there was no productive work to be done here; since *if there were reachable fruits, wouldn't academics be taking them*?

- Should I try my hand at becoming an entrepreneur? Whether or not it should be difficult to spot promising ideas in a scientific field, it certainly can't be easy to think up a profitable idea for a new startup. *Will I be able to find any good ideas that aren't already taken?*

- The effective altruism community is a network of philanthropists and researchers that try to find the very best ways to benefit others per dollar, in full generality. Where should effective altruism organizations like GiveWell expect to find low-hanging fruit—neglected interventions ripe with potential? *Where should they look to find things that our civilization isn't already doing about as well as can be done?*

When I think about problems like these, I use what feels to me like a natural generalization of the economic idea of efficient markets. The goal is to predict what kinds of efficiency we should expect to exist in realms beyond the marketplace, and what we can deduce from simple observations. For lack of a better term, I will call this kind of thinking *inadequacy analysis.*

Toward the end of this book, I'll try to refute an alternative viewpoint that is increasingly popular among some of my friends, one that I think is ill-founded. This

viewpoint is the one I've previously termed "modesty," and the message of modesty tends to be: "You can't expect to be able to do *X* that isn't usually done, since you could just be deluding yourself into thinking you're better than other people."

I'll open with a cherry-picked example that I think helps highlight the difference between these two viewpoints.

i.

I once wrote a report, "Intelligence Explosion Microeconomics," that called for an estimate of the economic growth rate in a fully developed country—that is, a country that is no longer able to improve productivity just by importing well-tested innovations. A footnote of the paper remarked that even though Japan was the country with the most advanced technology—e.g., their cellphones and virtual reality technology were five years ahead of the rest of the world's—I wasn't going to use Japan as my estimator for developed economic growth, because, as I saw it, Japan's monetary policy was utterly deranged.

Roughly, Japan's central bank wasn't creating enough money. I won't go into details here.

A friend of mine, and one of the most careful thinkers I know—let's call him "John"—made a comment on my draft to this effect:

> How do you claim to know this? I can think of plenty of other reasons why Japan could be in a slump: the country's shrinking and aging population, its low female workplace participation, its high levels of product market regulation, etc. It looks like you're venturing outside of your area of expertise to no good end.

"How do you claim to know this?" is a very reasonable question here. As John later elaborated, macroeconomics is an area where data sets tend to be thin and predictive performance tends to be poor. And John had previously observed me making contrarian claims where I'd turned out to be badly wrong, like endorsing Gary Taubes' theories about the causes of the obesity epidemic. More recently, John won money off of me by betting that AI performance on certain metrics would improve faster than I expected; John has a good track record when it comes to spotting my mistakes.

It's also easy to imagine reasons an observer might have been skeptical. I wasn't making up my critique of Japan myself; I was reading other economists and deciding that I trusted the ones who were saying that the Bank of Japan was doing it wrong… … Yet one would expect the governing board of the Bank of Japan to be composed of experienced economists with specialized monetary expertise. How likely is it that any outsider would be able to spot an obvious flaw in their policy? How likely is it that someone who isn't a professional economist (e.g., me) would be able to judge which economic critiques of the Bank of Japan were correct, or which critics were wise?

How likely is it that an entire country—one of the world's most advanced countries—would forego trillions of dollars of real economic growth because their monetary controllers—not politicians, but appointees from the professional elite—were doing something so wrong that even a non-professional could tell? How likely is it that a non-

professional could not just suspect that the Bank of Japan was doing something badly wrong, but be *confident* in that assessment?

Surely it would be more *realistic* to search for possible reasons why the Bank of Japan might not be as stupid as it seemed, as stupid as some econbloggers were claiming. Possibly Japan's aging population made growth impossible. Possibly Japan's massive outstanding government debt made even the slightest inflation too dangerous. Possibly we just aren't thinking of the complicated reasoning going into the Bank of Japan's decision.

Surely some *humility* is appropriate when criticizing the elite decision-makers governing the Bank of Japan. What if it's you, and not the professional economists making these decisions, who have failed to grasp the relevant economic considerations?

I'll refer to this genre of arguments as "modest epistemology."

In conversation, John clarified to me that he rejects this genre of arguments; but I hear these kinds of arguments fairly often. The head of an effective altruism organization once gave voice to what I would consider a good example of this mode of thinking:

> I find it helpful to admit to unpleasant facts that will necessarily be true in the abstract, in order to be more willing to acknowledge them in specific cases. For instance, I should expect a priori to be below average at half of things, and be 50% likely to be of below average talent overall; to know many people who I regard as better than me according to my values; to regularly make decisions that look silly ex post, and also ex ante; to be mistaken about issues on which there is expert disagreement about half of the time; to perform badly at many things I attempt for the first time; and so on.

The Dunning-Kruger effect shows that unskilled individuals often rate their own skill very highly. Specifically, although there does tend to be a correlation between how competent a person is and how competent they *guess* they are, this correlation is weaker than one might suppose. In the original study, people in the bottom two quartiles of actual test performance tended to think they did better than about 60% of test-takers, while people in the top two quartiles tended to think they did better than 70% of test-takers.

This suggests that a typical person's guesses about how they did on a test are evidence, but not particularly powerful evidence: the top quartile is underconfident in how well they did, and the bottom quartiles are highly overconfident.

Given all that, how can we gain much evidence from our belief that we are skilled? Wouldn't it be more prudent to remind ourselves of the base rate—the prior probability of 50% that we are below average?

Reasoning along similar lines, software developer Hal Finney has endorsed "abandoning personal judgment on most matters in favor of the majority view." Finney notes that the *average* person's opinions would be more accurate (on average) if they simply deferred to the most popular position on as many issues as they could. For this reason:

> I choose to adopt the view that in general, on most issues, the average opinion of humanity will be a better and less biased guide to the truth than my own judgment.

> […] I would suggest that although one might not always want to defer to the majority opinion, it should be the default position. Rather than starting with the assumption that one's own opinion is right, and then looking to see if the majority has good reasons for holding some other view, one should instead start off by following the majority opinion; and then only adopt a different view for good and convincing reasons. On most issues, the default of deferring to the majority will be the best approach. If we accept the principle that "extraordinary claims require extraordinary evidence", we should demand a high degree of justification for departing from the majority view. The mere fact that our own opinion seems sound would not be enough.[1]

In this way, Finney hopes to correct for overconfidence and egocentric biases.

Finney's view is an extreme case, but helps illustrate a pattern that I believe can be found in some more moderate and widely endorsed views. When I speak of "modesty," I have in mind a fairly diverse set of positions that rest on a similar set of arguments and motivations.

I once heard an Oxford effective altruism proponent crisply summarize what I take to be the central argument for this perspective: "You see that someone says *X*, which seems wrong, so you conclude their epistemic standards are bad. But they could just see that you say *Y*, which sounds wrong to them, and conclude your epistemic standards are bad."[2] On this line of thinking, you don't get any information about who has better epistemic standards merely by observing that someone disagrees with you. After all, the other side observes just the same fact of disagreement.

Applying this argument form to the Bank of Japan example: I receive little or no evidence just from observing that the Bank of Japan says "*X*" when I believe "not *X*." I also can't be getting strong evidence from any object-level impression I might have that I am unusually competent. So did my priors imply that I and I alone ought to have been born with awesome powers of discernment? (Modest people have posed this exact question to me on more than one occasion.)

It should go without saying that this isn't how I would explain my own reasoning. But if I reject arguments of the form, "We disagree, therefore I'm right and you're wrong," how can I claim to be correct on an economic question where I disagree with an institution as reputable as the Bank of Japan?

The other viewpoint, opposed to modesty—the view that I think is prescribed by normative epistemology (and also by more or less mainstream microeconomics)—requires a somewhat longer introduction.

ii.

By ancient tradition, every explanation of the Efficient Markets Hypothesis must open with the following joke:

> Two economists are walking along the street, and one says, "Hey, someone dropped a $20 bill!" and the other says, "Well, it can't be a real $20 bill because someone would have picked it up already."

Also by ancient tradition, the next step of the explanation is to remark that while it may make sense to pick up a $20 bill you see on a relatively deserted street, if you think you have spotted a $20 bill lying on the floor of Grand Central Station (the main subway nexus of New York City), and it has stayed there for several hours, then it probably *is* a fake $20 bill, or it has been glued to the ground.

In real life, when I asked a group of twenty relatively young people how many of them had ever found a $20 bill on the street, five raised their hands, and only one person had found a $20 bill on the street on two separate occasions. So the empirical truth about the joke is that while $20 bills on the street do exist, they're rare.

On the other hand, the implied policy is that if you do find a $20 bill on the street, you should go ahead and pick it up, because that does happen. It's not *that* rare. You certainly shouldn't start agonizing over whether it's too arrogant to believe that you have better eyesight than everyone else who has recently walked down the street.

On the other other hand, you *should* start agonizing about whether to trust your own mental processes if you think you've seen a $20 bill stay put for several hours on the floor of Grand Central Station. Especially if your explanation is that nobody else is eager for money.

Is there any other domain such that if we *think* we see an exploitable possibility, we should sooner doubt our own mental competence than trust the conclusion we reasoned our way to?

If I had to name the *single* epistemic feat at which modern human civilization is most adequate, the peak of all human power of estimation, I would unhesitatingly reply, "Short-term relative pricing of liquid financial assets, like the price of S&P 500 stocks relative to other S&P 500 stocks over the next three months." This is something into which human civilization puts an *actual effort*.

- Millions of dollars are offered to smart, conscientious people with physics PhDs to induce them to enter the field.
- These people are then offered huge additional payouts conditional on actual performance—especially outperformance relative to a baseline.[3]
- Large corporations form to specialize in narrow aspects of price-tuning.
- They have enormous computing clusters, vast historical datasets, and competent machine learning professionals.
- They receive repeated news of success or failure in a fast feedback loop.[4]
- The knowledge aggregation mechanism—namely, prices that equilibrate supply and demand for the financial asset—has proven to work beautifully, and acts to sum up the wisdom of all those highly motivated actors.
- An actor that spots a 1% systematic error in the aggregate estimate is rewarded with a billion dollars—in a process that also corrects the estimate.
- Barriers to entry are not zero (*you* can't get the loans to make a billion-dollar corrective trade), but there are thousands of diverse intelligent actors who are all individually allowed to spot errors, correct them, and be rewarded, with no central veto.

This is certainly not perfect, but it is *literally as good as it gets on modern-day Earth.*

I don't think I can beat the estimates produced by that process. I have no significant help to contribute to it. With study and effort I might become a decent hedge fundie and make a standard return. Theoretically, a liquid market should be just exploitable enough to pay competent professionals the same hourly rate as their next-best opportunity. I could potentially become one of those professionals, and earn standard hedge-fundie returns, but that's not the same as significantly improving on the market's efficiency. I'm not sure I expect a huge humanly accessible opportunity of that kind to *exist*, not in the thickly traded centers of the market. Somebody *really would* have taken it already! Our civilization *cares* about whether Microsoft stock will be priced at $37.70 or $37.75 tomorrow afternoon.

I can't predict a 5% move in Microsoft stock in the next two months, and *neither can you*. If your uncle tells an anecdote about how he tripled his investment in NetBet.com last year and he attributes this to his skill rather than luck, we know *immediately and out of hand* that he is wrong. Warren Buffett at the peak of his form couldn't reliably triple his money every year. If there is a strategy so simple that your uncle can understand it, which has apparently made him money—then we guess that there were just hidden risks built into the strategy, and that in another year or with less favorable events he would have lost half as much as he gained. Any other possibility would be the equivalent of a $20 bill staying on the floor of Grand Central Station for ten years while a horde of physics PhDs searched for it using naked eyes, microscopes, and machine learning.

In the thickly traded parts of the stock market, where the collective power of human civilization is truly at its strongest, I doff my hat, I put aside my pride and kneel in true humility to accept the market's beliefs as though they were my own, knowing that any impulse I feel to second-guess and every independent thought I have to argue otherwise is nothing but my own folly. If my perceptions suggest an exploitable opportunity, then my perceptions are far more likely mistaken than the markets. That is what it feels like to look upon a civilization doing something adequately.

The converse side of the efficient-markets perspective would have said this about the Bank of Japan:

CONVENTIONAL CYNICAL ECONOMIST: So, Eliezer, you think you know better than the Bank of Japan and many other central banks around the world, do you?

ELIEZER: Yep. Or rather, by reading econblogs, I believe myself to have identified which econbloggers know better, like Scott Sumner.

C.C.E.: Even though literally trillions of dollars of real value are at stake?

ELIEZER: Yep.

C.C.E.: How do you make money off this special knowledge of yours?

ELIEZER: I can't. The market also collectively knows that the Bank of Japan is pursuing a bad monetary policy and has priced Japanese equities accordingly. So even though I know the Bank of Japan's policy will make Japanese equities perform badly, that fact is already priced in; I can't expect to make money by short-selling Japanese equities.

C.C.E.: I see. So exactly who is it, on this theory of yours, that is being stupid and passing up a predictable payout?

ELIEZER: Nobody, of course! Only the Bank of Japan is allowed to control the trend line of the Japanese money supply, and the Bank of Japan's governors are not paid any bonuses when the Japanese economy does better. They don't get a million dollars in personal bonuses if the Japanese economy grows by a trillion dollars.

C.C.E.: So you can't make any money off knowing better individually, and nobody who has the actual power and authority to fix the problem would gain a personal financial benefit from fixing it? Then we're done! No anomalies here; this sounds like a perfectly normal state of affairs.

We don't usually expect to find $20 bills lying on the street, because even though people sometimes drop $20 bills, someone else will usually have a chance to pick up that $20 bill before we do.

We don't think we can predict 5% price changes in S&P 500 company stock prices over the next month, because we're competing against dozens of hedge fund managers with enormous supercomputers and physics PhDs, any one of whom could make millions or billions on the pricing error—and in doing so, correct that error.

We can expect it to be hard to come up with a truly good startup idea, and for even the best ideas to involve sweat and risk, because lots of other people are trying to think up good startup ideas. Though in this case we do have the advantage that we can pick our own battles, seek out *one* good idea that we think hasn't been done yet.

But the Bank of Japan is just one committee, and it's not possible for anyone else to step up and make a billion dollars in the course of correcting their error. Even if you think you know exactly what the Bank of Japan is doing wrong, you can't make a profit on that. At least some hedge-fund managers also know what the Bank of Japan is doing wrong, and the expected consequences are already priced into the market. Nor does this price movement fix the Bank of Japan's mistaken behavior. So to the extent the Bank of Japan has poor incentives or some other systematic dysfunction, their mistake can persist. As a consequence, when I read some econbloggers who I'd seen being right about empirical predictions before saying that Japan was being grotesquely silly, and the economic logic seemed to me to check out, as best I could follow it, I wasn't particularly reluctant to believe them. *Standard economic theory, generalized beyond the markets to other facets of society, did not seem to me to predict that the Bank of Japan must act wisely for the good of Japan.* It would be no surprise if they were competent, but also not much of a surprise if they were incompetent. And knowing this didn't help me either—I couldn't exploit the knowledge to make an excess profit myself—and this too wasn't a coincidence.

This kind of thinking can get quite a bit more complicated than the foregoing paragraphs might suggest. We have to ask why the government of Japan didn't put pressure on the Bank of Japan (answer: they did, but the Bank of Japan refused), and many other questions. You would need to consider a much larger model of the world, and bring in a lot more background theory, to be confident that you understood the overall situation with the Bank of Japan.

But even without that detailed analysis, in the epistemological background we have a completely different picture from the modest one. We have a picture of the world

where it is perfectly plausible for an econblogger to write up a good analysis of what the Bank of Japan is doing wrong, and for a sophisticated reader to reasonably agree that the analysis seems decisive, without a deep agonizing episode of Dunning-Kruger-inspired self-doubt playing any important role in the analysis.

iii.

When we critique a government, we don't usually get to see what would actually happen if the government took our advice. But in this one case, less than a month after my exchange with John, the Bank of Japan—under the new leadership of Haruhiko Kuroda, and under unprecedented pressure from recently elected Prime Minister Shinzo Abe, who included monetary policy in his campaign platform—embarked on an attempt to print huge amounts of money, with a stated goal of doubling the Japanese money supply.[5]

Immediately after, Japan experienced real GDP growth of 2.3%, where the previous trend was for falling RGDP. Their economy was operating that far under capacity due to lack of money.[6]

Now, on the modest view, this was the unfairest test imaginable. Out of all the times that I've ever suggested that a government's policy is suboptimal, the rare time a government tries my preferred alternative will select the most mainstream, highest-conventional-prestige policies I happen to advocate, and those are the very policy proposals that modesty is least likely to disapprove of.

Indeed, if John had looked further into the issue, he would have found (as I found while writing this) that Nobel laureates had also criticized Japan's monetary policy. He would have found that previous Japanese governments had also hinted to the Bank of Japan that they should print more money. The view from modesty looks at this state of affairs and says, "Hold up! You aren't so specially blessed as your priors would have you believe; other academics already know what you know! Civilization isn't so inadequate after all! This is how reasonable dissent from established institutions and experts operates in the real world: via opposition by other mainstream experts and institutions, not via the heroic effort of a lone economics blogger."

However helpful or unhelpful such remarks may be for guarding against inflated pride, however, they don't seem to refute (or even address) the central thesis of civilizational *inadequacy*, as I will define that term later. Roughly, the civilizational inadequacy thesis states that in situations where the central bank of a major developed democracy is carrying out a policy, and a number of highly regarded economists like Ben Bernanke have written papers about what that central bank is doing wrong, and there are widely accepted macroeconomic theories for understanding what that central bank is doing wrong, and the government of the country has tried to put pressure on the central bank to stop doing it wrong, and literally *trillions* of dollars in real wealth are at stake, then *the overall competence of human civilization* is such that we shouldn't be surprised to find the professional economists at the Bank of Japan doing it wrong.

We shouldn't even be surprised to find that a decision theorist without all that much background in economics can identify which econbloggers have correctly stated what

the Bank of Japan is doing wrong, or which simple improvements to their current policies would improve the situation.

iv.

It doesn't make much difference to my life whether I understand monetary policy better than, say, the European Central Bank, which as of late 2015 was repeating the same textbook mistake as the Bank of Japan and causing trillions of euros of damage to the European economy. Insofar as I have other European friends in countries like Italy, it might be important to them to know that Europe's economy is probably not going to get any better soon; or the knowledge might be relevant to predicting AI progress timelines to know whether Japan ran out of low-hanging technological fruit or just had bad monetary policy. But that's a rather distant relevance, and for most of my readers I would expect this issue to be even less relevant to their lives.

But you run into the same implicit background questions of inadequacy analysis when, for example, you're making health care decisions. Cherry-picking another anecdote: My wife has a severe case of Seasonal Affective Disorder. As of 2014, she'd tried sitting in front of a little lightbox for an hour per day, and it hadn't worked. SAD's effects were crippling enough for it to be worth our time to consider extreme options, like her spending time in South America during the winter months. And indeed, vacationing in Chile and receiving more exposure to actual sunlight *did* work, where lightboxes failed.

From my perspective, the obvious next thought was: "Empirically, dinky little lightboxes don't work. Empirically, the Sun does work. Next step: *more light*. Fill our house with more lumens than lightboxes provide." In short order, I had strung up sixty-five 60W-equivalent LED bulbs in the living room, and another sixty-five in her bedroom.

Ah, but should I assume that my civilization is being *opportunistic* about seeking out ways to cure SAD, and that if putting up 130 LED light bulbs often worked when lightboxes failed, *doctors would already know about that*? Should the fact that putting up 130 light bulbs isn't a well-known next step after lightboxes convince me that my bright idea is probably not a good idea, because if it were, everyone would already be doing it? Should I conclude from my inability to find any published studies on the Internet testing this question that there is some fatal flaw in my plan that I'm just not seeing?

We might call this argument "Chesterton's Absence of a Fence." The thought being: I shouldn't build a fence here, because if it were a good idea to have a fence here, someone would already have built it. The underlying question here is: How strongly should I expect that this extremely common medical problem has been thoroughly considered by my civilization, and that there's nothing new, effective, and unconventional that I can personally improvise?

Eyeballing this question, my off-the-cuff answer—based mostly on the impressions related to me by every friend of mine who has ever dealt with medicine on a research level—is that I wouldn't *necessarily* expect any medical researcher ever to have done a formal experiment on the first thought that popped into my mind for treating this extremely common depressive syndrome. Nor would I strongly expect the

intervention, if initial tests found it to be effective, to have received enough attention that I could Google it.

But this is just my personal take on the adequacy of 21st-century medical research. Should I be nervous that this line of thinking is just an excuse? Should I fret about the apparently high estimate of my own competence implied by my thinking that I could find an obvious-seeming way to remedy SAD when *trained doctors* aren't talking about it and I'm not a medical researcher? Am I going too far outside my own area of expertise and starting to think that I'm good at everything?

In practice, I didn't bother going through an agonizing fit of self-doubt along those lines. The systematic competence of human civilization with respect to treating mood disorders wasn't so apparent to me that I considered it a better use of resources to quietly drop the issue than to just lay down the ~$600 needed to test my suspicion. So I went ahead and ran the experiment. And as of early 2017, with two winters come and gone, Brienne seems to no longer have crippling SAD—though it took a *lot* of light bulbs, including light bulbs in her bedroom that had to be timed to go on at 7:30am before she woke up, to sustain the apparent cure.[7]

If you want to outperform—if you want to do anything not usually done—then you'll need to conceptually divide our civilization into areas of lower and greater competency. My view is that this is best done from a framework of incentives and the equilibria of those incentives—which is to say, from the standpoint of microeconomics. This is the main topic I'll cover here.

In the process, I will also make the case that modesty—the part of this process where you go into an agonizing fit of self-doubt—isn't actually helpful for figuring out when you might outperform some aspect of the equilibrium.

But one should initially present a positive agenda in discussions like these—saying first what you think is the correct epistemology, before inveighing against a position you think is wrong.

So without further ado, in the next chapter I shall present a very simple framework for inadequate equilibria.

Next chapter: **An Equilibrium of No Free Energy**.

The full book will be available November 16th. You can go to equilibriabook.com to pre-order the book, or sign up for notifications about new chapters and other developments.

---

1. See Finney, "Philosophical Majoritarianism." ↩

2. Note: They later said that I'd misunderstood their intent, so take this example with some grains of salt. ↩

3. This is why I specified *relative* prices: stock-trading professionals are usually graded on how well they do compared to the stock market, not compared to bonds. It's much less obvious that bonds in general are priced reasonably relative to stocks in general, though this is still being debated by economists. ↩

4. This is why I specified *near-term* pricing of liquid assets. ↩

5. That is, the Bank of Japan purchased huge numbers of bonds with newly created electronic money. ↩

6. See "How Japan Proved Printing Money Can Be A Great Idea" for a more recent update.

   For readers who are wondering, "Wait, how the heck can printing money possibly lead to real goods and services being created?" I suggest Googling "sticky wages" and possibly consulting Scott Sumner's history of the Great Depression, *The Midas Paradox*. ↩

7. Specifically, Brienne's symptoms were mostly cured in the winter of 2015, and partially cured in the winter of 2016, when she spent most of her time under fewer lights. Brienne reports that she suffered a lot less even in the more recent winter, and experienced no suicidal ideation, unlike in years prior to the light therapy.

   I'll be moderately surprised if this treatment works *reliably*, just because most things don't where depression is concerned; but I would predict that it works often enough to be worth trying for other people experiencing severe treatment-resistant SAD. ↩

Slack for your belief system

Follow-up to Zvi's post on Slack

You can have Slack in your life. But you can also have Slack in your belief system.

Initially, this seems like it might be bad.

Won't Slack result in a lack of precision? If I give myself Slack to believe in whatever, won't I just end up with a lot of wrong beliefs? Shouldn't I always be trying to *decrease* the amount of Slack in my beliefs, always striving to walk the narrow, true path?

Claims:

1. For some things, the only way to stumble upon the Truth is to have some Slack. In other words, having no Slack in your belief system can result in getting stuck at local optima.
2. Having Slack allows you to use fake frameworks in a way that isn't epistemically harmful.
3. If you are, in fact, just correct, I guess you should have zero Slack. But—just checking—are you ALSO correct about how you come to Know Things? If your way of coming to conclusions is even a little off, giving yourself zero Slack might be dangerous. (Having zero Slack in your meta process *multiplies* the problem of no-Slack to all downstream beliefs.)
4. I'm willing to make the more unbacked, harder-to-define claim that there exists no individual human alive who should have zero Slack in their beliefs, on the meta level. (In other words, no human has a truth-seeking process that will reliably get all the right answers.)

[I want to note that I fully believe I could be wrong about all four claims here, or thinking about this in the entirely wrong way. So fight me.]

Now, I'm going to specifically discuss Slack in one's *meta* process.

So, while I can apply the concept of Slack to individual beliefs themselves (aka "holding beliefs lightly"), I am applying the concept more to the question of "How do I come to know/understand anything or or call a thing true?"

So, I'm not discussing examples of "I believe X, with more or less Slack." I'm discussing the difference between, "Doing a bunch of studies is the only way to know things" (less Slack) vs. "Doing a bunch of studies is how I currently come to know things, but I'm open to other ways" (more Slack).

The less Slack there is in your process for forming beliefs, the more constraints you have to abide before being able to claim you've come to understand something.

Examples of such constraints include:

- I only buy it if it has had at least one peer-reviewed RCT.
- This framework seems like it'll lead to confirmation bias, so I will ignore it.
- If it involves politics or tribalism or status, it can't have any truth to it.
- If it's self-contradictory / paradoxical, it has to be one way or the other.

- I can't imagine this being true or useful because my gut reaction to it is negative.
- I don't feel anything about it, so it must be meaningless.
- This doesn't conform to my narrative or worldview. In fact it's offensive to consider, so I won't.
- If I thought this, it would likely result in harm to myself or others, so I can't think it.
- It's only true if I can prove it.
- It's only worth considering if it's been tested empirically.
- I should discard models that aren't made of gears.

Note that sometimes, it is good to have such constraints, at least for now.

Not everyone can interact with facts, claims, and beliefs without some harm to their epistemics. In fact, most people cannot, I claim. (And further, I believe this to be one of the most important problems in rationality.)

That said, I see a lot of people's orientations as:

"My belief-forming process says this thing isn't true, and in fact this entire class of thing is likely false and not worth digging into. You seem to be actively engaging with [class of thing] and claiming there is truth in it. That seems highly dubious—there is something wrong with your belief-forming process."

This is a reasonable stance to take.

After all, lots of things aren't worth digging into. And lots of people have bad truth-seeking processes. Theirs may very well be worse than yours; you don't have to consider something just because it's in front of you.

But if you notice yourself unwilling to engage with [entire class of thing]... to me this indicates something is suboptimal.

Over time, it seems good to aim for being able to engage with more classes of things, rather than fewer.

If something is politically charged, yes, your beliefs are at risk, and you may be better off avoiding the topic altogether. But—wouldn't it be nice, if one day, you could wade through the mire of politics and come out the other side, clean? Epistemics in tact? Even better, you come out the other side having realized new truths about the world?

I guess if I'm going to be totally honest, the reason I am saying this is because I feel annoyed when people dismiss entire [classes of thing] for reasons like, "That part of the territory is really swampy and dangerous! Going in there is bad, and you're probably compromised."

At least *some* of the time, the thing that is going on is the person just figured out how to navigate swamps.

But instead, I feel like the person lacks Slack in their belief-forming process and is also trying to enforce this lack of Slack onto others.

From the inside, I imagine this feels like, "No one can navigate swamps, and anyone who says they are is probably *terribly* mistaken or naive about how truth-seeking works, so I should inform them of the danger."

From the inside, Slack will feel incorrect or potentially dangerous. Without constraints, the person may feel like they'll go off the rails—maybe they'll even end up believing in *gasp* horoscopes or *gasp* the existence of a Judeo-Christian God.

My greatest fear is not having false beliefs. My greatest fear is getting trapped into a particular definition of truth-seeking, such that I permanently end up with many false beliefs or large gaps in my map.

The two things I do to avoid this are:

a) Learn more skills for navigating tricky territories. For example, one of the skills is noticing a belief that's in my mind because it would be beneficial for me to believe it, i.e. it makes me feel good in a certain way or I expect good things to happen as a result—say, it'd make a person like me more if I believed it. This likely requires a fair amount of introspective capacity.

b) Be open to the idea that other people have truth-seeking methods that I don't. That they're seeing entire swaths of reality I can't see. Be curious about that, and try to learn more. Develop taste around this. Maintain some Slack, so I don't become myopic.

There's No Fire Alarm for Artificial General Intelligence

What is the function of a fire alarm?

One might think that the function of a fire alarm is to provide you with important evidence about a fire existing, allowing you to change your policy accordingly and exit the building.

In the classic experiment by Latane and Darley in 1968, eight groups of three students each were asked to fill out a questionnaire in a room that shortly after began filling up with smoke. Five out of the eight groups didn't react or report the smoke, even as it became dense enough to make them start coughing. Subsequent manipulations showed that a lone student will respond 75% of the time; while a student accompanied by two actors told to feign apathy will respond only 10% of the time. This and other experiments seemed to pin down that what's happening is pluralistic ignorance. We don't want to look panicky by being afraid of what isn't an emergency, so we try to look calm while glancing out of the corners of our eyes to see how others are reacting, but of course they are also trying to look calm.

(I've read a number of replications and variations on this research, and the effect size is blatant. I would not expect this to be one of the results that dies to the replication crisis, and I haven't yet heard about the replication crisis touching it. But we have to put a maybe-not marker on everything now.)

A fire alarm creates common knowledge, in the you-know-I-know sense, that there is a fire; after which it is socially safe to react. When the fire alarm goes off, you know that everyone else knows there is a fire, you know you won't lose face if you proceed to exit the building.

The fire alarm doesn't tell us with certainty that a fire is there. In fact, I can't recall one time in my life when, exiting a building on a fire alarm, there was an actual fire. Really, a fire alarm is *weaker* evidence of fire than smoke coming from under a door.

But the fire alarm tells us that it's socially okay to react to the fire. It promises us with certainty that we won't be embarrassed if we now proceed to exit in an orderly fashion.

It seems to me that this is one of the cases where people have mistaken beliefs about what they believe, like when somebody loudly endorsing their city's team to win the big game will back down as soon as asked to bet. They haven't consciously distinguished the rewarding exhilaration of shouting that the team will win, from the feeling of anticipating the team will win.

When people look at the smoke coming from under the door, I think they think their uncertain wobbling feeling comes from not assigning the fire a high-enough probability of really being there, and that they're reluctant to act for fear of wasting effort and time. If so, I think they're interpreting their own feelings mistakenly. If that was so, they'd get the same wobbly feeling on hearing the fire alarm, or even more so, because fire alarms correlate to fire less than does smoke coming from under a door. The uncertain wobbling feeling comes from the worry that others believe differently, not the worry that the fire isn't there. The reluctance to act is the

reluctance to be seen looking foolish, not the reluctance to waste effort. That's why the student alone in the room does something about the fire 75% of the time, and why people have no trouble reacting to the much weaker evidence presented by fire alarms.

* * *

It's now and then proposed that we ought to start reacting later to the issues of Artificial General Intelligence (background here), because, it is said, we are so far away from it that it just isn't possible to do productive work on it today.

(For direct argument about there being things doable today, see: Soares and Fallenstein (2014/2017); Amodei, Olah, Steinhardt, Christiano, Schulman, and Mané (2016); or Taylor, Yudkowsky, LaVictoire, and Critch (2016).)

(If none of those papers existed or if you were an AI researcher who'd read them but thought they were all garbage, and you wished you could work on alignment but knew of nothing you could do, the wise next step would be to sit down and spend two hours by the clock sincerely trying to think of possible approaches. Preferably without self-sabotage that makes sure you don't come up with anything plausible; as might happen if, hypothetically speaking, you would actually find it much more comfortable to believe there was nothing you ought to be working on today, because e.g. then you could work on other things that interested you more.)

(But never mind.)

So if AGI seems far-ish away, and you think the conclusion licensed by this is that you can't do any productive work on AGI alignment yet, then the implicit alternative strategy on offer is: Wait for some unspecified future event that tells us AGI is coming near; and *then* we'll all know that it's okay to start working on AGI alignment.

This seems to me to be wrong on a number of grounds. Here are some of them.

One: As Stuart Russell observed, if you get radio signals from space and spot a spaceship there with your telescopes and you know the aliens are landing in thirty years, you still start thinking about that today.

You're not like, "Meh, that's thirty years off, whatever." You certainly don't casually say "Well, there's nothing we can do until they're closer." Not without spending two hours, or at least five minutes by the clock, brainstorming about whether there is anything you ought to be starting now.

If you said the aliens were coming in thirty years and you were therefore going to do nothing today... well, if these were more effective times, somebody would ask for a schedule of what you thought ought to be done, starting when, how long before the aliens arrive. If you didn't have that schedule ready, they'd know that you weren't operating according to a worked table of timed responses, but just procrastinating and doing nothing; and they'd correctly infer that you probably hadn't searched very hard for things that could be done today.

In Bryan Caplan's terms, anyone who seems quite casual about the fact that "nothing can be done now to prepare" about the aliens is missing a mood; they should be much more alarmed at not being able to think of any way to prepare. And maybe ask if somebody else has come up with any ideas? But never mind.

Two: History shows that for the general public, and even for scientists not in a key inner circle, and even for scientists *in* that key circle, it is very often the case that key technological developments still seem decades away, five years before they show up.

In 1901, two years before helping build the first heavier-than-air flyer, Wilbur Wright told his brother that powered flight was fifty years away.

In 1939, three years before he personally oversaw the first critical chain reaction in a pile of uranium bricks, Enrico Fermi voiced 90% confidence that it was impossible to use uranium to sustain a fission chain reaction. I believe Fermi also said a year after that, aka two years before the denouement, that *if* net power from fission was even possible (as he then granted some greater plausibility) then it would be fifty years off; but for this I neglected to keep the citation.

And of course if you're not the Wright Brothers or Enrico Fermi, you will be even more surprised. Most of the world learned that atomic weapons were now a thing when they woke up to the headlines about Hiroshima. There were esteemed intellectuals saying four years *after* the Wright Flyer that heavier-than-air flight was impossible, because knowledge propagated more slowly back then.

Were there events that, in hindsight, today, we can see as signs that heavier-than-air flight or nuclear energy were nearing? Sure, but if you go back and read the actual newspapers from that time and see what people actually said about it then, you'll see that they did not know that these were signs, or that they were very uncertain that these might be signs. Some playing the part of Excited Futurists proclaimed that big changes were imminent, I expect, and others playing the part of Sober Scientists tried to pour cold water on all that childish enthusiasm; I expect that part was more or less exactly the same decades earlier. If somewhere in that din was a superforecaster who said "decades" when it was decades and "5 years" when it was five, good luck noticing them amid all the noise. More likely, the superforecasters were the ones who said "Could be tomorrow, could be decades" both when the big development was a day away and when it was decades away.

One of the major modes by which hindsight bias makes us feel that the past was more predictable than anyone was actually able to predict at the time, is that in hindsight we know what we ought to notice, and we fixate on only one thought as to what each piece of evidence indicates. If you look at what people actually say at the time, historically, they've usually got no clue what's about to happen three months before it happens, because they don't know which signs are which.

I mean, you *could* say the words “AGI is 50 years away” and have those words happen to be true. People were also saying that powered flight was decades away when it was in fact decades away, and those people happened to be right. The problem is that everything looks the same to you either way, if you are actually living history instead of reading about it afterwards.

It's not that whenever somebody says "fifty years" the thing always happens in two years. It's that this confident prediction of things being far away corresponds to an epistemic state about the technology that feels the same way internally until you are very very close to the big development. It's the epistemic state of "Well, I don't see how to do the thing" and sometimes you say that fifty years off from the big development, and sometimes you say it two years away, and sometimes you say it while the Wright Flyer is flying somewhere out of your sight.

Three: Progress is driven by peak knowledge, not average knowledge.

If Fermi and the Wrights couldn't see it coming three years out, imagine how hard it must be for anyone else to see it.

If you're not at the global peak of knowledge of how to do the thing, and looped in on all the progress being made at what will turn out to be the leading project, you aren't going to be able to see of your own knowledge *at all* that the big development is imminent. Unless you are very good at perspective-taking in a way that wasn't necessary in a hunter-gatherer tribe, and very good at realizing that other people may know techniques and ideas of which you have no inkling even that you do not know them. If you don't consciously compensate for the lessons of history in this regard; then you will promptly say the decades-off thing. Fermi wasn't still thinking that net nuclear energy was impossible or decades away by the time he got to 3 months before he built the first pile, because at that point Fermi was looped in on everything and saw how to do it. But anyone not looped in probably still felt like it was fifty years away while the actual pile was fizzing away in a squash court at the University of Chicago.

People don't seem to automatically compensate for the fact that the timing of the big development is a function of the peak knowledge in the field, a threshold touched by the people who know the most and have the best ideas; while they themselves have average knowledge; and therefore what they themselves know is not strong evidence about when the big development happens. I think they aren't thinking about that at all, and they just eyeball it using their own sense of difficulty. If they are thinking anything more deliberate and reflective than that, and incorporating real work into correcting for the factors that might bias their lenses, they haven't bothered writing down their reasoning anywhere I can read it.

To know that AGI is decades away, we would need enough understanding of AGI to know what pieces of the puzzle are missing, and how hard these pieces are to obtain; and that kind of insight is unlikely to be available until the puzzle is complete. Which is also to say that to anyone outside the leading edge, the puzzle will look more incomplete than it looks on the edge. That project may publish their theories in advance of proving them, although I hope not. But there are unproven theories now too.

And again, that's not to say that people saying "fifty years" is a certain sign that something is happening in a squash court; they were saying “fifty years” sixty years ago too. It's saying that anyone who thinks technological *timelines* are actually forecastable, in advance, by people who are not looped in to the leading project's progress reports and who don't share all the best ideas about exactly how to do the thing and how much effort is required for that, is learning the wrong lesson from history. In particular, from reading history books that neatly lay out lines of progress and their visible signs that we all know *now* were important and evidential. It's sometimes possible to say useful conditional things about the consequences of the big development whenever it happens, but it’s rarely possible to make confident predictions about the *timing* of those developments, beyond a one- or two-year horizon. And if you are one of the rare people who can call the timing, if people like that even exist, nobody else knows to pay attention to you and not to the Excited Futurists or Sober Skeptics.

Four: The future uses different tools, and can therefore easily do things that are very hard now, or do with difficulty things that are impossible now.

Why do we know that AGI is decades away? In popular articles penned by heads of AI research labs and the like, there are typically three prominent reasons given:

(A) The author does not know how to build AGI using present technology. The author does not know where to start.

(B) The author thinks it is really very hard to do the impressive things that modern AI technology does, they have to slave long hours over a hot GPU farm tweaking hyperparameters to get it done. They think that the public does not appreciate how hard it is to get anything done right now, and is panicking prematurely because the public thinks anyone can just fire up Tensorflow and build a robotic car.

(C) The author spends a lot of time interacting with AI systems and therefore is able to personally appreciate all the ways in which they are still stupid and lack common sense.

We've now considered some aspects of argument A. Let's consider argument B for a moment.

Suppose I say: "It is now possible for one comp-sci grad to do in a week anything that N+ years ago the research community could do with neural networks *at all*." How large is N?

I got some answers to this on Twitter from people whose credentials I don't know, but the most common answer was five, which sounds about right to me based on my own acquaintance with machine learning. (Though obviously not as a literal universal, because reality is never that neat.) If you could do something in 2012 period, you can probably do it fairly straightforwardly with modern GPUs, Tensorflow, Xavier initialization, batch normalization, ReLUs, and Adam or RMSprop or just stochastic gradient descent with momentum. The modern techniques are just that much better. To be sure, there are things we can't do now with just those simple methods, things that require tons more work, but those things were not possible at all in 2012.

In machine learning, when you can do something at all, you are probably at most a few years away from being able to do it easily using the future's much superior tools. From this standpoint, argument B, "You don't understand how hard it is to do what we do," is something of a non-sequitur when it comes to timing.

Statement B sounds to me like the same sentiment voiced by Rutherford in 1933 when he called net energy from atomic fission "moonshine". If you were a nuclear physicist in 1933 then you had to split all your atoms by hand, by bombarding them with other particles, and it was a laborious business. If somebody talked about getting net energy from atoms, maybe it made you feel that you were unappreciated, that people thought your job was easy.

But of course this will always be the lived experience for AI engineers on serious frontier projects. You don't get paid big bucks to do what a grad student can do in a week (unless you're working for a bureaucracy with no clue about AI; but that's not Google or FB). Your personal experience will *always* be that what you are paid to spend months doing is difficult. A change in this personal experience is therefore not something you can use as a fire alarm.

Those playing the part of wiser sober skeptical scientists would obviously agree in the abstract that our tools will improve; but in the popular articles they pen, they just talk about the painstaking difficulty of this year's tools. I think that when they're in that

mode they are not even trying to forecast what the tools will be like in 5 years; they haven't written down any such arguments as part of the articles I've read. I think that when they tell you that AGI is decades off, they are literally giving an estimate of how long it feels to them like it would take to build AGI using their current tools and knowledge. Which is why they emphasize how hard it is to stir the heap of linear algebra until it spits out good answers; I think they are not imagining, at all, into how this experience may change over considerably less than fifty years. If they've explicitly considered the bias of estimating future tech timelines based on their present subjective sense of difficulty, and tried to compensate for that bias, they haven't written that reasoning down anywhere I've read it. Nor have I ever heard of that forecasting method giving good results historically.

Five: Okay, let's be blunt here. I don't think most of the discourse about AGI being far away (*or* that it's near) is being generated by models of future progress in machine learning. I don't think we're looking at wrong models; I think we're looking at no models.

I was once at a conference where there was a panel full of famous AI luminaries, and most of the luminaries were nodding and agreeing with each other that of course AGI was very far off, except for two famous AI luminaries who stayed quiet and let others take the microphone.

I got up in Q&A and said, "Okay, you've all told us that progress won't be all that fast. But let's be more concrete and specific. I'd like to know what's the *least* impressive accomplishment that you are very confident *cannot* be done in the next two years."

There was a silence.

Eventually, two people on the panel ventured replies, spoken in a rather more tentative tone than they'd been using to pronounce that AGI was decades out. They named "A robot puts away the dishes from a dishwasher without breaking them", and Winograd schemas. Specifically, "I feel quite confident that the Winograd schemas--where we recently had a result that was in the 50, 60% range--in the next two years, we will not get 80, 90% on that regardless of the techniques people use."

A few months after that panel, there was unexpectedly a big breakthrough on Winograd schemas. The breakthrough didn't crack 80%, so three cheers for wide credibility intervals with error margin, but I expect the predictor might be feeling slightly more nervous now with one year left to go. (I don't think it was the breakthrough I remember reading about, but Rob turned up this paper as an example of one that could have been submitted at most 44 days after the above conference and gets up to 70%.)

But that's not the point. The point is the silence that fell after my question, and that eventually I only got two replies, spoken in tentative tones. When I asked for concrete feats that were impossible in the next two years, I think that that's when the luminaries on that panel switched to trying to build a mental model of future progress in machine learning, asking themselves what they could or couldn't predict, what they knew or didn't know. And to their credit, most of them did know their profession well enough to realize that forecasting future boundaries around a rapidly moving field is actually *really hard*, that nobody knows what will appear on arXiv next month, and that they needed to put wide credibility intervals with very generous upper bounds on how much progress might take place twenty-four months' worth of arXiv papers later.

(Also, Demis Hassabis was present, so they all knew that if they named something insufficiently impossible, Demis would have DeepMind go and do it.)

The question I asked was in a completely different genre from the panel discussion, requiring a mental context switch: the assembled luminaries actually had to try to consult their rough, scarce-formed intuitive models of progress in machine learning and figure out what future experiences, if any, their model of the field definitely prohibited within a two-year time horizon. Instead of, well, emitting socially desirable verbal behavior meant to kill that darned hype about AGI and get some predictable applause from the audience.

I'll be blunt: I don't think the confident long-termism has been thought out at all. If your model has the extraordinary power to say what will be impossible in ten years after another one hundred and twenty months of arXiv papers, then you ought to be able to say much weaker things that are impossible in two years, and you should have those predictions queued up and ready to go rather than falling into nervous silence after being asked.

In reality, the two-year problem is hard and the ten-year problem is laughably hard. The future is hard to predict in general, our predictive grasp on a rapidly changing and advancing field of science and engineering is very weak indeed, and it doesn't permit narrow credible intervals on what can't be done.

Grace et al. (2017) surveyed the predictions of 352 presenters at ICML and NIPS 2015. Respondents' aggregate forecast was that the proposition "all occupations are fully automatable" (in the sense that "for any occupation, machines could be built to carry out the task better and more cheaply than human workers") will not reach 50% probability until 121 years hence. Except that a randomized subset of respondents were instead asked the slightly different question of "when unaided machines can accomplish every task better and more cheaply than human workers", and in this case held that this was 50% likely to occur within 44 years.

That's what happens when you ask people to produce an estimate they can't estimate, and there's a social sense of what the desirable verbal behavior is supposed to be.

* * *

When I observe that there's no fire alarm for AGI, I'm not saying that there's no possible equivalent of smoke appearing from under a door.

What I'm saying rather is that the smoke under the door is always going to be arguable; it is not going to be a clear and undeniable and absolute sign of fire; and so there is never going to be a fire alarm producing common knowledge that action is now due and socially acceptable.

There's an old trope saying that as soon as something is actually done, it ceases to be called AI. People who work in AI and are in a broad sense pro-accelerationist and techno-enthusiast, what you might call the Kurzweilian camp (of which I am not a member), will sometimes rail against this as unfairness in judgment, as moving goalposts.

This overlooks a real and important phenomenon of adverse selection against AI accomplishments: If you can do something impressive-sounding with AI in 1974, then that is because that thing turned out to be doable in some cheap cheaty way, not

because 1974 was so amazingly great at AI. We are uncertain about how much cognitive effort it takes to perform tasks, and how easy it is to cheat at them, and the first "impressive" tasks to be accomplished will be those where we were most wrong about how much effort was required. There was a time when some people thought that a computer winning the world chess championship would require progress in the direction of AGI, and that this would count as a sign that AGI was getting closer. When Deep Blue beat Kasparov in 1997, in a Bayesian sense we did learn something about progress in AI, but we also learned something about chess being easy. Considering the techniques used to construct Deep Blue, most of what we learned was "It is surprisingly possible to play chess without easy-to-generalize techniques" and not much "A surprising amount of progress has been made toward AGI."

Was AlphaGo smoke under the door, a sign of AGI in 10 years or less? People had previously given Go as an example of What You See Before The End.

Looking over the paper describing AlphaGo's architecture, it seemed to me that we *were* mostly learning that available AI techniques were likely to go further towards generality than expected, rather than about Go being surprisingly easy to achieve with fairly narrow and ad-hoc approaches. Not that the method scales to AGI, obviously; but AlphaGo did look like a product of *relatively* general insights and techniques being turned on the special case of Go, in a way that Deep Blue wasn’t. I also updated significantly on "The general learning capabilities of the human cortical algorithm are less impressive, less difficult to capture with a ton of gradient descent and a zillion GPUs, than I thought," because if there were anywhere we expected an impressive hard-to-match highly-natural-selected but-still-general cortical algorithm to come into play, it would be in humans playing Go.

Maybe if we'd seen a thousand Earths undergoing similar events, we'd gather the statistics and find that a computer winning the planetary Go championship is a reliable ten-year-harbinger of AGI. But I don't actually know that. Neither do you. Certainly, anyone can publicly argue that we just learned Go was easier to achieve with strictly narrow techniques than expected, as was true many times in the past. There's no possible sign short of actual AGI, no case of smoke from under the door, for which we know that this is definitely serious fire and now AGI is 10, 5, or 2 years away. Let alone a sign where we know everyone else will believe it.

And in any case, multiple leading scientists in machine learning have already published articles telling us their criterion for a fire alarm. They will believe Artificial General Intelligence is imminent:

(A) When they personally see how to construct AGI using their current tools. This is what they are always saying is not currently true in order to castigate the folly of those who think AGI might be near.

(B) When their personal jobs do not give them a sense of everything being difficult. This, they are at pains to say, is a key piece of knowledge not possessed by the ignorant layfolk who think AGI might be near, who only believe that because they have never stayed up until 2AM trying to get a generative adversarial network to stabilize.

(C) When they are very impressed by how smart their AI is relative to a human being in respects that still feel magical to them; as opposed to the parts they do know how to engineer, which no longer seem magical to them; aka the AI seeming pretty smart in interaction and conversation; aka the AI actually being an AGI already.

So there isn't going to be a fire alarm. Period.

There is never going to be a time before the end when you can look around nervously, and see that it is now clearly common knowledge that you can talk about AGI being imminent, and take action and exit the building in an orderly fashion, without fear of looking stupid or frightened.

* * *

So far as I can presently estimate, now that we've had AlphaGo and a couple of other maybe/maybe-not shots across the bow, and seen a huge explosion of effort invested into machine learning and an enormous flood of papers, we are probably going to occupy our present epistemic state until very near the end.

By saying we're probably going to be in roughly this epistemic state until almost the end, I *don't* mean to say we know that AGI is imminent, or that there won't be important new breakthroughs in AI in the intervening time. I mean that it's hard to guess how many further insights are needed for AGI, or how long it will take to reach those insights. After the next breakthrough, we still won't know how many more breakthroughs are needed, leaving us in pretty much the same epistemic state as before. Whatever discoveries and milestones come next, it will probably continue to be hard to guess how many further insights are needed, and timelines will continue to be similarly murky. Maybe researcher enthusiasm and funding will rise further, and we'll be able to say that timelines are shortening; or maybe we'll hit another AI winter, and we'll know that's a sign indicating that things will take longer than they would otherwise; but we still won't know *how long.*

At some point we might see a sudden flood of arXiv papers in which really interesting and fundamental and scary cognitive challenges seem to be getting done at an increasing pace. Whereupon, as this flood accelerates, even some who imagine themselves sober and skeptical will be unnerved to the point that they venture that perhaps AGI is only 15 years away now, maybe, possibly. The signs might become so blatant, very soon before the end, that people start thinking it is socially acceptable to say that maybe AGI is 10 years off. Though the signs would have to be pretty darned blatant, if they're to overcome the social barrier posed by luminaries who are estimating arrival times to AGI using their personal knowledge and personal difficulties, as well as all the historical bad feelings about AI winters caused by hype.

But even if it becomes socially acceptable to say that AGI is 15 years out, in those last couple of years or months, I would still expect there to be disagreement. There will still be others protesting that, as much as associative memory and human-equivalent cerebellar coordination (or whatever) are now solved problems, they still don't know how to construct AGI. They will note that there are no AIs writing computer science papers, or holding a truly sensible conversation with a human, and castigate the senseless alarmism of those who talk as if we already knew how to do that. They will explain that foolish laypeople don't realize how much pain and tweaking it takes to get the current systems to work. (Although those modern methods can easily do almost anything that was possible in 2017, and any grad student knows how to roll a stable GAN on the first try using the tf.unsupervised module in Tensorflow 5.3.1.)

When all the pieces are ready and in place, lacking only the last piece to be assembled by the very peak of knowledge and creativity across the whole world, it will still seem to the average ML person that AGI is an enormous challenge looming in the distance, because they still won't personally know how to construct an AGI system.

Prestigious heads of major AI research groups will still be writing articles decrying the folly of fretting about the total destruction of all Earthly life and all future value it could have achieved, and saying that we should not let this distract us from *real, respectable concerns* like loan-approval systems accidentally absorbing human biases.

Of course, the future is very hard to predict in detail. It's so hard that not only do I confess my own inability, I make the far stronger positive statement that nobody else can do it either. The “flood of groundbreaking arXiv papers” scenario is one way things could maybe possibly go, but it's an implausibly specific scenario that I made up for the sake of concreteness. It's certainly not based on my extensive experience watching other Earthlike civilizations develop AGI. I do put a significant chunk of probability mass on "There's not much sign visible outside a Manhattan Project until Hiroshima," because that scenario is simple. Anything more complex is just one more story full of burdensome details that aren't likely to all be true.

But no matter how the details play out, I do predict in a very general sense that there will be no fire alarm that is not an actual running AGI--no unmistakable sign before then that everyone knows and agrees on, that lets people act without feeling nervous about whether they're worrying too early. That's just not how the history of technology has usually played out in much simpler cases like flight and nuclear engineering, let alone a case like this one where all the signs and models are disputed. We already know enough about the uncertainty and low quality of discussion surrounding this topic to be able to say with confidence that there will be no unarguable socially accepted sign of AGI arriving 10 years, 5 years, or 2 years beforehand. If there’s any general social panic it will be by coincidence, based on terrible reasoning, uncorrelated with real timelines except by total coincidence, set off by a Hollywood movie, and focused on relatively trivial dangers.

It's no coincidence that nobody has given any actual account of such a fire alarm, and argued convincingly about how much time it means we have left, and what projects we should only then start. If anyone does write that proposal, the next person to write one will say something completely different. And probably neither of them will succeed at convincing me that they know anything prophetic about timelines, or that they've identified any sensible angle of attack that is (a) worth pursuing at all and (b) not worth starting to work on right now.

* * *

It seems to me that the decision to delay all action until a nebulous totally unspecified future alarm goes off, implies an order of recklessness great enough that the law of continued failure comes into play.

The law of continued failure is the rule that says that if your country is incompetent enough to use a plaintext 9-numeric-digit password on all of your bank accounts and credit applications, your country is not competent enough to correct course after the next disaster in which a hundred million passwords are revealed. A civilization competent enough to correct course in response to that prod, to react to it the way you'd want them to react, is competent enough not to make the mistake in the first place. When a system fails massively and obviously, rather than subtly and at the very edges of competence, the next prod is not going to cause the system to suddenly snap into doing things intelligently.

The law of continued failure is especially important to keep in mind when you are dealing with big powerful systems or high-status people that you might feel nervous about derogating, because you may be tempted to say, "Well, it's flawed now, but as soon as a future prod comes along, everything will snap into place and everything will be all right." The systems about which this fond hope is actually warranted look like they are mostly doing all the important things right already, and only failing in one or two steps of cognition. The fond hope is almost never warranted when a person or organization or government or social subsystem is currently falling massively short.

The folly required to ignore the prospect of aliens landing in thirty years is already great enough that the other flawed elements of the debate should come as no surprise.

And with all of that going wrong simultaneously today, we should predict that the same system and incentives won't produce correct outputs after receiving an uncertain sign that maybe the aliens are landing in five years instead. The law of continued failure suggests that if existing authorities failed in enough different ways at once to think that it makes sense to try to derail a conversation about existential risk by saying the real problem is the security on self-driving cars, the default expectation is that they will still be saying silly things later.

People who make large numbers of simultaneous mistakes don’t generally have all of the incorrect thoughts subconsciously labeled as "incorrect" in their heads. Even when motivated, they can't suddenly flip to skillfully executing all-correct reasoning steps instead. Yes, we have various experiments showing that monetary incentives can reduce overconfidence and political bias, but (a) that's reduction rather than elimination, (b) it's with extremely clear short-term direct incentives, not the nebulous and politicizable incentive of "a lot being at stake", and (c) that doesn't mean a switch is flipping all the way to "carry out complicated correct reasoning". If someone's brain contains a switch that can flip to enable complicated correct reasoning at all, it's got enough internal precision and skill to think mostly-correct thoughts now instead of later--at least to the degree that some conservatism and double-checking gets built into examining the conclusions that people know will get them killed if they’re wrong about them.

There is no sign and portent, no threshold crossed, that suddenly causes people to wake up and start doing things systematically correctly. People who can react that competently to any sign at all, let alone a less-than-perfectly-certain not-totally-agreed item of evidence that is *likely* a wakeup call, have probably already done the timebinding thing. They've already imagined the future sign coming, and gone ahead and thought sensible thoughts earlier, like Stuart Russell saying, "If you know the aliens are landing in thirty years, it's still a big deal now."

* * *

Back in the funding-starved early days of what is now MIRI, I learned that people who donated last year were likely to donate this year, and people who last year were planning to donate "next year" would quite often this year be planning to donate "next year". Of course there were genuine transitions from zero to one; everything that happens needs to happen for a first time. There were college students who said "later" and gave nothing for a long time in a genuinely strategically wise way, and went on to get nice jobs and start donating. But I also learned well that, like many cheap and easy solaces, saying the word "later" is addictive; and that this luxury is available to the rich as well as the poor.

I don't expect it to be any different with AGI alignment work. People who are trying to get what grasp they can on the alignment problem will, in the next year, be doing a little (or a lot) better with whatever they grasped in the previous year (plus, yes, any general-field advances that have taken place in the meantime). People who want to defer that until after there's a better understanding of AI and AGI will, after the next year's worth of advancements in AI and AGI, want to defer work until a better future understanding of AI and AGI.

Some people really *want* alignment to *get done* and are therefore *now* trying to wrack their brains about how to get something like a reinforcement learner to reliably identify a utility function over particular elements in a model of the causal environment instead of a sensory reward term or defeat the seeming tautologicalness of updated (non-)deference. Others would rather be working on other things, and will therefore declare that there is no work that can possibly be done today, *not* spending two hours quietly thinking about it first before making that declaration. And this will not change tomorrow, unless perhaps tomorrow is when we wake up to some interesting newspaper headlines, and probably not even then. The luxury of saying "later" is not available only to the truly poor-in-available-options.

After a while, I started telling effective altruists in college: "If you're planning to earn-to-give later, then for now, give around $5 every three months. And never give exactly the same amount twice in a row, or give to the same organization twice in a row, so that you practice the mental habit of re-evaluating causes and re-evaluating your donation amounts on a regular basis. *Don't* learn the mental habit of just always saying 'later'."

Similarly, if somebody was *actually* going to work on AGI alignment "later", I'd tell them to, every six months, spend a couple of hours coming up with the best current scheme they can devise for aligning AGI and doing useful work on that scheme. Assuming, if they must, that AGI were somehow done with technology resembling current technology. And publishing their best-current-scheme-that-isn't-good-enough, at least in the sense of posting it to Facebook; so that they will have a sense of embarrassment about naming a scheme that does not look like somebody actually spent two hours trying to think of the best bad approach.

There are things we'll better understand about AI in the future, and things we'll learn that might give us more confidence that particular research approaches will be relevant to AGI. There may be more future sociological developments akin to Nick Bostrom publishing *Superintelligence*, Elon Musk tweeting about it and thereby heaving a rock through the Overton Window, or more respectable luminaries like Stuart Russell openly coming on board. The future will hold more AlphaGo-like events to publicly and privately highlight new ground-level advances in ML technique; and it may somehow be that this does *not* leave us in the same epistemic state as having already seen AlphaGo and GANs and the like. It could happen! I can't see exactly how, but the future does have the capacity to pull surprises in that regard.

But before waiting on that surprise, you should ask whether your uncertainty about AGI timelines is really uncertainty at all. If it feels to you that guessing AGI might have a 50% probability in N years is not enough knowledge to act upon, if that feels scarily uncertain and you want to wait for more evidence before making any decisions... then ask yourself how you'd feel if you believed the probability was 50% in N years, and everyone else on Earth also believed it was 50% in N years, and everyone believed it was right and proper to carry out policy P when AGI has a 50% probability of arriving in N years. If that visualization feels very different, then any nervous "uncertainty" you

feel about doing P is not really about whether AGI takes much longer than N years to arrive.

And you are almost surely going to be stuck with that feeling of "uncertainty" no matter how close AGI gets; because no matter how close AGI gets, whatever signs appear will almost surely not produce common, share, agreed-on public knowledge that AGI has a 50% chance of arriving in N years, nor any agreement that it is therefore right and proper to react by doing P.

And if all that did become common knowledge, then P is unlikely to still be a neglected intervention, or AI alignment a neglected issue; so you will have waited until sadly late to help.

But far more likely is that the common knowledge just isn't going to be there, and so it will always feel nervously "uncertain" to consider acting.

You can either act despite that, or not act. Not act until it's too late to help much, in the best case; not act at all until after it's essentially over, in the average case.

I don't think it's wise to wait on an unspecified epistemic miracle to change how we feel. In all probability, you're going to be in this mental state for a while - including any nervous-feeling "uncertainty". If you handle this mental state by saying "later", that general policy is not likely to have good results for Earth.

* * *

Further resources:

- MIRI's research guide (https://intelligence.org/research-guide/) and forum (https://agentfoundations.org)
- FLI's collection of introductory resources
- CHAI's alignment bibliography at http://humancompatible.ai/bibliography
- 80,000 Hours' AI job postings on https://80000hours.org/job-board/
- The Open Philanthropy Project's AI fellowship and general call for research proposals
- My brain-dumps on AI alignment
- If you're arriving here for the first time, my long-standing work on rationality, and CFAR's workshops
- And some general tips from Ray Arnold for effective altruists considering AI alignment as a cause area.

Four Scopes Of Advice

There's a very common failure mode people fall into where they'll ask for 'advice on doing something', receive excellent advice, and fail to follow it. For a long time this was mysterious to me. Then a friend provided a possible explanation that completely changed how I look at it. As I explained to my friend, the problem isn't that the person giving advice isn't trustworthy, quite often the person asking wants the advice and trusts the opinion of the person giving it. So why don't they follow it? My friend hypothesized that people let their identity get mixed up in how they wanted to do the thing, and then can't bring themselves to do it another way. Essentially his hypothesis is that they're being asked to change too much. This seems plausible enough, but it got me thinking more broadly about what scope of advice people are looking for when they ask.

To start we can imagine four modes of planning, divided thusly:

- **Mission** - What you are trying to accomplish.
- **Strategic Planning** - What broad goals you intend to satisfy to get there.
- **Tactical Planning** - Concrete near term objectives which will let you satisfy the strategic goals.
- **Operational Planning** - The absolute lowest levels of getting the work done, who will do what, what needs to be done to make tactics work, etc.

The hierarchy of willingness to take advice then is basically a mirror image of this.

- **Unwilling** - The zeroth level. The one receiving advice is willing to change nothing based on what counsel they are given.
- **Operational Advice** ("Use this kind of bolt.") - The one receiving advice would like to hear suggestions on how to accomplish the task they've already set for themselves, but aren't particularly interested in hearing what tasks make sense in the service of what goals. This is probably what most people asking for advice actually want.
- **Tactical Advice** ("The car should have four wheels.") - Given a set of broad goals, the one receiving advice is open for suggestions on how they should go about trying to accomplish them. This might mean for example that significant deviations from the original plan are allowed as long as they better serve the goals which the action is going towards. Most advice I give is on this level whether it's asked for or not.
- **Strategic Advice** ("You should build a car.") - Very close to being the most open to radically plan changing advice. Here the one receiving advice is willing to accept that the goals they've decided on to pursue their mission are flawed, or perhaps not the best goals they could set for accomplishing the mission. This kind of advice is usually only solicited at the outset of a project, at least for a while. Once a project is in motion the inertia to change these becomes much higher, as a consequence people can persist in doing stupid things for essentially rational reasons even after it's been laboriously pointed out to them why the thing is dumb.

- **Mission Advice** ("You should build a high speed transportation machine.") - Here the one receiving advice is open to the idea that the thing they are trying to accomplish, may not even be the right thing to go after at all. This is probably the rarest kind of advice to be followed, and the rarest to be solicited once a project is in motion. Examples might include certain kinds of Effective Altruist activism that tries to convince people to quit their mediocre job and become an investment banker so they can donate the money to charity. Or maybe if in the course of trying to accomplish strategic goals an organization falls so far below what it hoped to accomplish that it begins to make more sense to 'pivot'.

The takeaway for you dear reader is that you should try to be cognizant of what kind of advice you're looking for. To get better advice it may help to explicitly communicate your preference to the person you're soliciting from. You're liable to make your peers quite angry if they give you solid strategic advice and you persist in the same basic inoptimal tactics towards your goals.

(This post was originally published at https://namespace.obormot.net/Main/FourScopesOfAdvice . Special thanks to Oliver for helping me polish it up.)

Against naming things, and so on

Recent discussion on naming concepts mostly focuses on arguments in favor, noting only a few caveats, as LW user Conor Moreton in Why and How to Name Things:

> What you lose by the proliferation of jargon is ease-of-entry and cross-cultural intelligibility and hard-drive space in the brains of people trying to track all of it.

I think you lose more than that, so I'll try to name [*sic*] a few more reasons your caution might run deeper.

1. The nomothetic fallacy

Diagnosis isn't a cure, but it can bring a sense of relief untethered from prognosis. You might still have the same symptoms that still need the same treatment, but it feels like a large part of the problem has been solved. This is, I'd argue, a decent chunk of why knowing about biases can hurt you. Don't get complacent just because you have a name for something.

2. Weaponized rationality

Another big chunk of "knowing about biases can hurt you" comes from wielding concepts not in introspection but against others. (Like many of the considerations below, taking this as compelling is a general argument against much of the rationality project. But proliferating jargon, to my eye, makes most of the problems here relatively worse for external use as compared to internal use.)

3. Being wrong

Sometimes your analysis of the thing you're trying to crystallize just isn't very good. Your analogy doesn't play out the way you think, the pattern doesn't actually exist, you're not carving reality at the joints, your framing could be better, it's an inappropriate level of abstraction anywhere you'd actually want to use it. But, hey, the name stuck, so you can communicate all that in just a word!

4. Lossy compression

Maybe you got it right this time, but the name doesn't capture everything. In practice, nobody's going to remember your entire blog post every time someone utters the title. I hope you didn't need that nuance. But at least this one went viral!

5. Thinking on the page

From Conor's post:

> When you define a concept rigorously and clearly, you almost always learn new things from playing around with their edges and trying to get them to interface

with other rigorous, clear concepts.

And if you don't define it rigorously and clearly—as is the case for pretty much everything happening here—if you play directly with the compressed concepts and string them together that way, oops, you just proved that 2+2=5. You lost track of the actual stuff the name referred to, the rules for manipulating it while preserving truth. You thought you were post-rigor when there isn't even a rigorous stage. You were thinking on the page. [Oh, look, links to named chunked concepts. What are they doing here, of all places?]

Combining frameworks and making analogies between your concepts is a good way of generating new ideas, but it's far from rigorous and if anything encourages sloppiness if you're not careful to recognize that it's less formal reasoning and more another way of pointing to where the new things to learn might be.

6. Illusion of transparency

People will assume they know what you mean. You'll assume you know what they mean in reply—that is, what a coincidence, that they know what you mean. They don't.

7. Lifting the pot by one handle

The trouble with rationalist skills is that the opposite of every rationalist skill is also a rationalist skill.

–komponisto

In the comments to Epistemic Learned Helplessness (now there's a name, eh?), komponisto writes:

> We have the Inside View, and the Outside View. Overconfidence is a problem, but so is underconfidence. You're supposed to listen to the tiniest note of mental discord, yet sometimes it's necessary to shut loud mental voices out. And while knowing the standard catalog of biases is obviously crucial for the aspiring rationalist, it can also hurt you. Et cetera, et cetera.

Is it lotus eating or self-care? Should you reverse any advice you hear or not? Remembering a blogpost title isn't going to tell you which situation you're in; it's at best a handle for noticing you're in a situation where you might want to course correct following deeper analysis. But implicit in the assignment of handles is a claim that you ought to adjust more than you do in a certain direction. And you don't want achieving a balance of opposite mistakes to depend on the relative catchiness of named concepts. When you lift the pot by one handle, the soup spills out.

(This isn't, on its own, an argument against giving your ideas catchy names, so much as an argument for spending plenty of time discussing trade-offs and how to do the necessary deeper analysis. But recall that it's still mostly the handle that gets remembered.)

8. Reification

You'll tend to treat your named ideas as more concrete than they are, as meaning one particular thing to everyone, as having agency or causal powers, as unchanging and non-reframeable, as universal and context-independent, as territory, as binary, as objective.

(I wonder if this is related to a lot of our apparent confusion about double crux—people (including me) questioning its status as The Technique, while Duncan keeps trying to clarify how it's not about executing an algorithm, it's about internalizing what generates the algorithm; it's not designed for use in earnest, it's a pedagogical tool for practicing mental habits; it's not about Double Crux, it's about double-cruxiness.)

9. Rounding error

Conor again:

> [What you gain by the proliferation of jargon includes] being embedded in a culture where people are diligently seeking out and popularizing such distinctions makes a given individual far more likely to pay attention to subtle distinctions themselves, accelerating the process of cultural accumulation of nuance and detail.

Or, well, the opposite of that, where you round everything off to ingroup-approved jargon.

***

You can't stop titling your blog posts, but don't force it, or you'll end up with virality tracking the wrong things (more so than usual, anyway). In particular, maybe don't optimize so much towards things like catchiness, metaphorical weight, uniqueness, and communicability—even for the sake of rapid, rigorous, distinction-rich, high-level discussion (or perhaps especially for that sake, since this kind of chunking is either unnecessary or destructive for all of these but speed)—over choosing words that prepare people to read and understand what you say.

(And even as far as speed of communication, what's weird is that the rationalist writing stereotype is absurdly prolix—it feels like the compression isn't happening where it matters, just where it feels powerful for hiding complex abstraction in a minimally intelligible way.)

You shouldn't stop trying to draw more distinctions, but maybe you can do that without encouraging one particular way of pointing to your distinction to reify and metastasize.

You won't stop sloppily compressing and combining complex ideas, but maybe worry more about unpacking things, giving examples, checking understanding, and asking questions.

The Problematic Third Person Perspective

[Epistemic status: I now endorse this again. Michael pointed out a possibility for downside risk with losing mathematical ability, which initially made me update away from the view here. However, some experience noticing what it is like to make certain kinds of mathematical progress made me return to the view presented here. Maybe don't take this post as inspiration to engage in extreme rejection of objectivity.]

There are a number of conversational norms based on the idea of an imaginary impartial observer who needs to be convinced. It's the adversarial courtroom model of conversation. Better norms, such as common crux, can be established by recognizing that a conversation is taking place between two people.

Burden-of-proof is one of these problematic ideas. The idea that there is some kind of standard which would put the burden on one person or another would only make sense if there were a judge to convince. If anything, it would be better to say the burden of proof is on both people in any argument, in the sense that they are responsible for conveying their own views to the other person. If burden-of-proof is about establishing that they "should" give in to your position, it accomplishes nothing; you need to convince *them* of that, not yourself. If burden-of-proof is about establishing that you don't have to believe them until they say more... well, that was true anyway, but perhaps speaks to a lack of curiosity on your part.

More generally, this external-judge intuition promotes the bad model that there are objective standards of logic which must be adhered to in a debate. There *are* epistemic standards which it is *good* to adhere to, including logic and notions of probabilistic evidence. But, if the other person has different standards, then you have to either work with them or discuss the differences. There's a failure mode of the overly rationalistic where you just get angry that *their* arguments are illogical and they're not accepting *your* perfectly-formatted arguments, so you try to get them to bow down to your standards by force of will. (The same failure mode applies to treating definitions as objective standards which must be adhered to.) What good does it do to continue arguing with them via standards you already know differ from theirs? Try to understand and engage with their real reasons rather than replacing them with imaginary things.

Actually, it's even worse than this, because you don't know your own standards of evidence completely. So, the imaginary impartial judge is also interfering with your ability to get in touch with your real reasons, what you really think, and what might sway you one way or the other. If your mental motion is to reach for justifications which the impartial judge would accept, you are rationalizing rather than finding your true rejection. You have to realize that you're using standards of evidence that you yourself don't fully understand, and live in that world -- otherwise you rob yourself of the ability to improve your tools.

This happens in two ways, that I can think of.

- Maybe your explicit standards are good, but not perfect. You notice beliefs that are not up to your standards, and you drop them reflexively. This might be a good idea most of the time, but there are two things wrong with the policy. First,

you might have dropped a good belief. You could have done better by checking which you trusted more in this instance: the beliefs, or your standards of belief. Second, you've missed an opportunity to improve your explicit standards. You could have explored your reasons for believing what you did, and compared them to your explicit standards for belief.
- Maybe you don't notice the difference between your explicit standards and the way you actually arrive at your beliefs. You assume implicitly that if you believe something strongly, it's because there are strong reasons of the sort you endorse. This is especially likely if the beliefs pattern-match to the sort of thing your standards endorse; for example, being very sciency. As a result, you miss an opportunity to notice that you're rationalizing something. You would have done better to first look for the reasons you *really* believed the thing, and then check whether they meet your explicit standards and whether the belief still seems worth endorsing.

So far, I've argued that the imaginary judge creates problems in two domains: navigating disagreements with other people, and navigating your own epistemic standards. I'll note a third domain where the judge seems problematic: judging your own actions and decisions. Many people use an imaginary judge to guide their actions. This leads to pitfalls such as moral self-licensing, in which doing good things gives you a license to do more bad things (setting up a budget makes you feel good enough about your finances that you can go on a spending spree, eating a salad for lunch makes you more likely to treat yourself with ice cream after work, etc). Getting rid of the internal judge is an instance of Nate's Replacing Guilt, and carries similar risks: if you're currently using the internal judge for a bunch of important things, you have to either make sure you replace it with other working strategies, or be OK with kicking those things to the roadside (at least temporarily).

Similarly with the other two categories I mentioned. Noticing the dysfunctions of the imaginary-judge perspective should not make you immediately remove it; invoke Chesterton's Fence. However, I would encourage you to experiment with removing the imaginary third person from your conversations, and seeing what you do when you remind yourself that there's no one looking over your shoulder in your private mental life. I think this relates to a larger ontological shift which Val was also pointing toward in In Praise of Fake Frameworks. There is no third-person perspective. There is no view from nowhere. This isn't a rejection of reductionism, but a reminder that we haven't finished yet. This isn't a rejection of the principles of rationality, but a reminder that we are created already in motion, and there is no argument so persuasive it would move a rock.

And, more basically, it is a reminder that the map is not the territory, because humans confuse the two by default. The picture in your head isn't what's there to be seen. Putting pieces of your judgement inside an imaginary impartial judge doesn't automatically make it true. Perhaps it does really make it more trustworthy -- you "promote" your better heuristics by wrapping them up inside the judge, giving them authority over the rest. But, this system has its problems. It can create perverse incentives on the other parts of your mind, to please the judge in ways that let them get away with what they want. It can make you blind to other ways of being. It can make you *think* you've avoided map-territory confusion once and for all -- "See? It's written right there on my soul: DO NOT CONFUSE MAP AND TERRITORY. It is simply something I don't do." -- while really passing the responsibility to a special part of your map which is now almost *always* confused for the territory.

So, laugh at the judge a little. Look out for your real reasons for thinking and doing things. Notice whether your arguments seem tailored to convince your judge rather than the person in front of you. See where it leads you.

I Can Tolerate Anything Except Factual Inaccuracies

This is a linkpost for http://greyenlightenment.com/i-can-tolerate-anything-except-factual-inaccuracies/

A wonderful post by greyenlightenment that touches on contrarian and intellectualism signalling. It mentions the dilemma between agreeing with the broad thrust of a piece, and agreeing with factual claims of the piece. We are suggested to consider not criticising a piece when we agree with the message but find little factual inaccuracies —a norm against nitpicking so to speak.

I suspect a norm against nitpicking would destroy a chesterton fence and lead down a slippery slope into anti-intellectualism and greater irrationality.
As Julia Galef says:

> Not caring about validity of an argument, as long as conclusion is true ~=
>
> Not caring about due process, as long as guilty guy is convicted

I think the same criticism appears to relaxing the norm against factual inaccuracies.

If we stop caring about whether the facts of the matter are very correct, then what next? I suspect the long term consequences of such a norm to be detrimental.

If it leads to a reduction in the quantity of articles I'll otherwise agree with (because the authors wanted to be as accurate as possible), then that's a trade off I would gladly accept.

I do recognise that I am a contrarian and love to signal intellectualism—for what it's worth.

What are your thoughts?

Writing That Provokes Comments

Epistemic Effort: Thought about it for a year. Solicited feedback. Checked my last few posts' comment count to make sure I wasn't *obviously* *wrong.*

A thing that happens to me, and perhaps to you:

Someone writes a beautiful essay that I agree with, that sheds new light on something important.

I don't have anything really to say about it. I don't want to just say "I agree!". So instead of commenting, I give it an upvote and move on.

This feels bad for a few reasons:

- I like commenting.
- I like *getting* comments when I write things that (I hope!) are insightful, beautiful and true. It's a stronger signal that people care.
- Comments correlate with something *staying in the public sphere of attention*. A highly upvoted post eventually fades behind newer upvoted posts. But a post with lots of comments keeps people paying attention (with new people constantly checking in to see what the hubbub is about)
- I don't trust (as a reader or a writer) that people who read a post, give it an upvote, and move on, are really *learning* anything. I think that talking through an a new concept and figuring out how to apply is where much of the learning happens.

I've been impressed with how much quality writing has been going on on LW2.0 so far. There has been *some* but not *as much* commenting as I'd like.

I've gotten a sense of what inspires interesting, meaty discussion.

Unfortunately, most of it seems... kinda bad?

Things That Get People To Comment

1. Be Wrong - It has been said: if google fails you, the fastest way to get a question answered is to post a *wrong answer* on reddit. This will result in a lot of flood of people explaining things to you.

2. Be Controversial - Even better, post something that *some* people think are wrong. Then you get a bunch of people commenting to correct you, and then other people who disagree correcting *them!* The arguments perpetuate themselves from there. You won't even have to do any commenting work yourself to keep it going!

[*BTW, these are observations, not recommendations. This list is optimized to answer the question "what causes comments" not "how to make the world better."]*

3. Write About Things People Feel Qualified to Have Opinions On - If you write a post on machine learning, and post it somewhere where nobody really understands machine learning, it doesn't matter if you're wrong or controversial! Nobody will understand enough to care, or feel confident enough to argue. Some considerations:

- It's not necessary for people to be qualified. They just need to feel like they are.
- If you write more informally (or in informal forums), people feel more entitled to respond.
- You can either tailor your topic to an existing audience, or p*roactively try to get an* existing audience who understands your weird niche topic to read your post.

4. Invoke Social Reality - People pay more attention when you're talking about social norms, or about changing coalitions of people, or arguing that some people are Bad and Wrong. This is for two reasons:

- Social Reality is powerful and scary. A person's sense of social safety is one of the most important things to them. People like to know who is Bad and Wrong so that they can be on the other side. People like making sure that if social norms changing, they are changing in ways they understand and like (so that nobody later decides they are Bad and Wrong).
- Social Reality almost always has something confusing and dumb going on that needs fixing, that people think is worth thinking about.
- People *understand* Social Reality. Or, they think they do. (See #3)
- Social Reality is often controversial! (See #2)

5. Be So Inspiring That People Create Entire Fandoms of Your Work - This worked for Eliezer and arguably Scott. It can probably be broken down into smaller steps. It's pretty hard though. And a bunch of people *trying* but failing to do this can be annoying. (*I've* tried/failed to do this sometimes)

...

And then there's...

6. Leave People With An Unsolved Problem That They Care About - This is related to "they feel qualified to have opinions", with the followup step of "there is actual useful thinking they can contribute to, either to solve *your* problem, or to apply your idea to solve *their* problems."

Things I've Noticed Myself Doing

Since comments are socially validating, I've noticed a tendency for me to end up writing:

- Facebook posts, where people feel a lower barrier to entry. (If the shortform section of LessWrong were up, I might do that instead)
- Unfinished thoughts, where there's a good chance that I'm wrong about a few things (but not all things, and not wrong *on purpose to be provocative* which would feel skeezy*),* and where there's still an unsolved problem that people will feel qualified to help out figure out.
- Posts engaging with social norms (which people feel excited to weigh in on and/or afraid not to)
- Posts engaging with personal habits that people can easily apply to their own life.

This doesn't all seem *bad,* necessarily. But I've noticed other people that seem to be doing similar things. I've also noticed some people who tried to get people to talk

about important things, and failed, and gradually resorted to writing more provocative things to get people to pay attention (which succeeded!).

It seems like a rationality community warped by those incentives isn't going to accomplish the things it needs to.

So, some open problems I'm thinking about, which maybe are relevant to you:

- I'd like feel **incentivized to research things I don't understand** as much (which I don't expect other people to understand as much either), to expand my (and our collective) domains of expertise.
- Insofar as people do end up writing the sorts of posts listed above, I think it'd be good if people **thought more consciously and carefully** about which tools they're employing. #6 at the very least seemed fine, and some of the others seem fine in some contexts.
- I'd like to **learn how to be a better commenter**, on posts that *don't* go out of their way to make it easy to comment. I have a sense that if I took the step of *actually stopping to think for a half-hour* about possible ramifications of a given post, I could probably think of something worth saying, and that it might get easier with time. (I've been thinking about that for the past week or two, but keep end up spending that time mostly writing my own posts, or engaging with other commenters who did more heavy lifting of initiating discussion)
- I'd like *people who have important things to say* to **be able to trust that people will listen**, without falling into an attentional arms race that leads inevitably to BuzzFeed. But right now *I* have trouble paying attention to things that are important but non-drama-laden, so I can't reasonably expect people to trust in that.

That's all I got for now.

Multidimensional signaling

What do you infer about a person who has ugly clothing? Probably that they have poor taste (in clothes, or subcultures). But it could also be that they are too poor to improve their wardrobe. Or can’t be bothered.

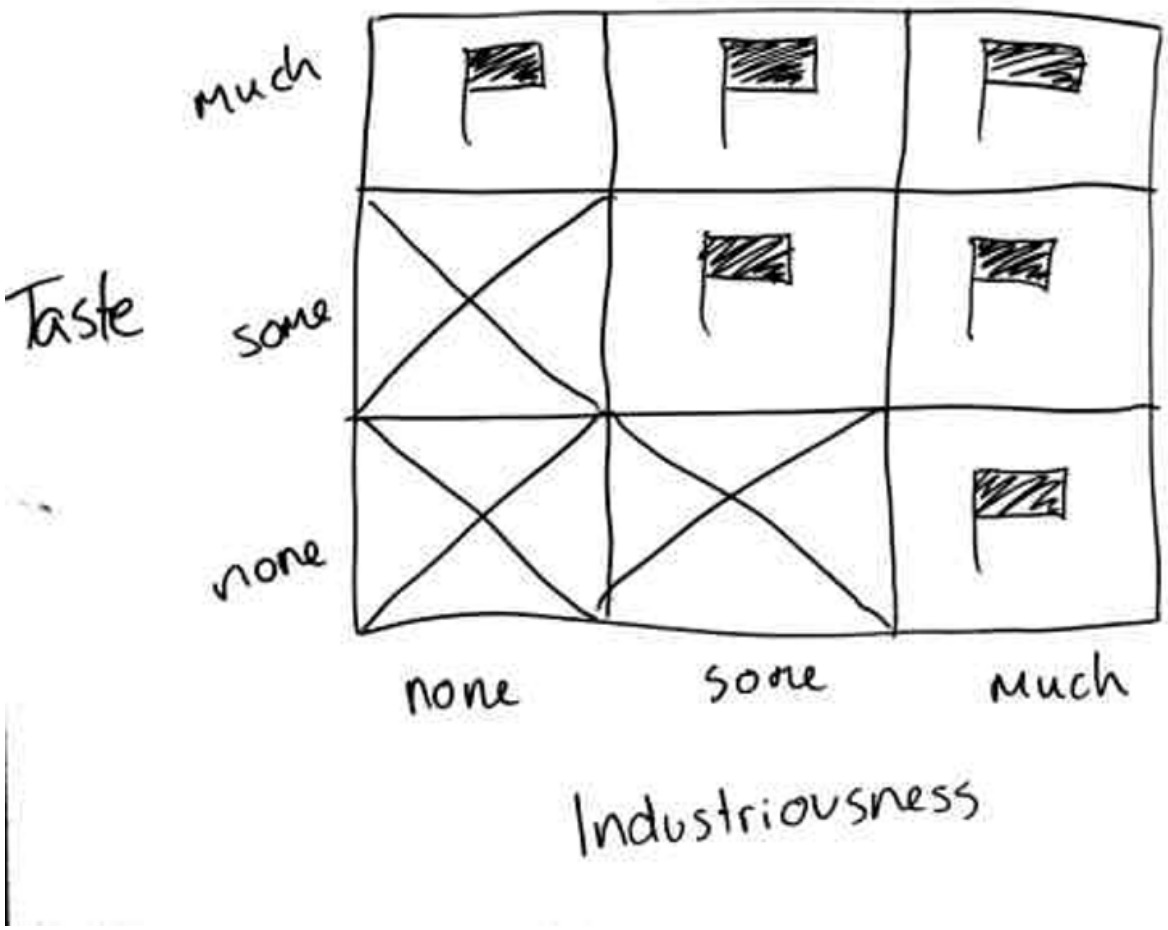
What about someone with poor grades? The obvious inference is that they aren’t so capable at the subject, but it may again be that they can’t be bothered, or that they have more urgent things to do with their time.

And someone who makes clever jokes? Probably that they are smart and naturally funny, but if they had more time or effort to spend on this, it probably helped.

For all kinds of traits that people might try to signal with their behavior, someone can send a better signal if they have more money or time or self-control. Even when the main signal being sent is not usually thought to be about any of those things.

The reason this interests me: if signals often divide the population into ‘better or richer’ vs. ‘worse or poorer’, I wonder if this would cause us to imagine that being rich is associated with being better, even if the two were entirely independent. (And similarly for wealth in other general-use resources, like self-control and time).

In a simple case, suppose there are just people with pretty clothes (who are both rich and have good taste) and people with ugly clothes (who either have bad taste, or lack resources or will). Then do observers come to think of ‘rich good taste’ type and a ‘poor bad taste’ type? Or do they pay more attention to the actual structure of the space, and know for instance that this doesn’t mean learning that someone really has bad taste actually means they are probably poorer.



Note that I'm not merely suggesting that a person with more wealth can send signals to look like they are better—that much is clear. I'm suggesting that at a population level, if the wealthier people can't be distinguished from the better people on some axis, then observers may come to think that the two are associated in general, even if they are not at all.

If so, this would be important, because it would apply in a huge range of cases of signaling. So that the properties of poverty and weak-willedness and such would appear to us to be much worse than they really were.

Windows Resource Repository

One tradition on LessWrong has been to have repository for neat tricks for various occasions. A lot of them are listed in the Repository repository.

In this iteration I want to focus on which useful Windows programs we have installed on our computers. If anybody has written interesting and useful scripts that they want to share, that's also welcomed.

That's a Thing!

Lauren Lee responded on facebook to my post about locating yourself as an instance of a class:

> Riffing on this LW post by Abram Demski (https://www.lesserwrong.com/…/placing-yourself-as-an-instan…)
>
> In particular this paragraph:
>
> << For example, if a person is trying to save money but sees a doodad they'd like to buy, the fool reason as follows: "It's just this one purchase. The amount of money isn't very consequential to my overall budget. I can just save a little more in other ways and I'll meet my target." The wise person reasons as follows: "If I make this purchase now, I will similarly allow myself to make exceptions to my money-saving rule later, until the exception becomes the rule and I spend all my money. So, even though the amount of money here isn't so large, I prefer to follow a general policy of saving, which implies saving in this particular case." A very wise person may reason a bit more cleverly: "I can make impulse purchases if they pass a high bar, such that I actually only let a few dollars of unplanned spending past the bar every week on average. How rare is it that a purchase opportunity costing this much is at least this appealing?" **does a quick check and usually doesn't buy the thing, but sometimes does, when it is worth it** >>
>
> The above is a rough outline for one set of moves you can make, but there are ways you can enhance the decision process even further.
>
> One way you can generally increase your "wisdom" is to notice what specific environmental cues are /relevant/ to WHY the impulse occurred in the first place. I think these cues are initially not obvious, at least in my experience.
>
> To illustrate, say I walk into a random store.
>
> What factors determine the likelihood that I'm going to want to impulsively buy a thing?
>
> There are times when I walk into a store without a plan, and I can be confident I'm going to walk out of it without having purchased anything. Other times, I feel more like I will have ended up purchasing something. What is giving me these clues?
>
> The situation of me being in a grocery store while I'm hungry is a different /class/ of situation from me being in a grocery store while I'm full. And so I want to treat these classes as somewhat separate when making decisions.
>
> Somehow, I figured out at some point that this was a different class. Maybe by noticing a pattern over time that I bought more stuff when I was hungry in grocery stores vs. full.
>
> The question is how to figure out one's own patterns / cues very generally, so they cover a wide range of situations.
>
> I suspect there's a bunch of tools for this, but one I made up just now:

> At the moment of environmental switch (e.g. I just walked into a store, I just saw something cool, I just noticed a desire arise), make a prediction about your future on the order of ~15-60 minutes. (This will hopefully have the effect of you checking in with your current state, taking inventory / capturing all the variables in that moment. With those variables—even before walking through the store further or considering a decision—you should be able to predict the outcome. That is my claim anyway.)
>
> In most situations, my own behavior shouldn't surprise me. I have most of the relevant information about what my behavior will be. Humans are mostly just TAP machines. If you know what relevant Triggers to be paying attention to, you should be able to make accurate predictions about the output Actions.
>
> And yes, you can change the output Actions, but mostly by changing what Triggers you pay attention to / how to weight them, not by trying to reprogram your Actions in response to the exact same state.
>
> (If you've ever found yourself disappointed in yourself or surprised by yourself, you should try to figure out what the Trigger was that caused it.)

To react to a thing, some part of you needs to recognize that it is a thing.

In Retroactive Readmission of Evidence, Conor says:

> Eventually, though, the pattern makes itself clear, and catches my mental eye, and I'm like *Gah, this thing*! It's only on iteration number seven that I realize that, a) that thing is really annoying, and b) it's happened six times before, too.

In Why and How to Name Things, Conor discusses how a proliferation of named things broadens what we can think about.

In Outside View as the Main Debiasing Technique, I discuss how simply knowing about a bias can allow you to wake up when the bias is about to happen, and course-correct. Grognor and I expressed a similar sentiment in A List of Nuances. Grognor writes about why lumping errors are easier to make than splitting errors.

All of these things assume that potential categories come from somewhere, but don't say much about where they come from. You can take outside view all you want and never get anywhere if you have the wrong categories. But how do we get better categories?

Lauren Lee suggests we make predictions and watch for what drives our behavior. I agree with her that there's probably a bunch of advice for this. But, I suspect none of it will stick very well until we have a good name for the problem. What do we call the problem of cutting the world at its joints?

Lots of bad names come to mind. Reification. Ontology. Class formation, concept formation, category formation. Object segmentation. Factoring the world. For now, I think all of them are trumped by "*That's a thing!*".

AlphaGo Zero and the Foom Debate

AlphaGo Zero uses 4 TPUs, is built entirely out of neural nets with no handcrafted features, doesn't pretrain against expert games or anything else human, reaches a superhuman level after 3 days of self-play, and is the strongest version of AlphaGo yet.

The architecture has been simplified. Previous AlphaGo had a policy net that predicted good plays, and a value net that evaluated positions, both feeding into lookahead using MCTS (random probability-weighted plays out to the end of a game). AlphaGo Zero has one neural net that selects moves and this net is trained by Paul Christiano-style capability amplification, playing out games against itself to learn new probabilities for winning moves.

As others have also remarked, this seems to me to be an element of evidence that favors the Yudkowskian position over the Hansonian position in my and Robin Hanson's AI-foom debate.

As I recall and as I understood:

- Hanson doubted that what he calls "architecture" is much of a big deal, compared to (Hanson said) elements like cumulative domain knowledge, or special-purpose components built by specialized companies in what he expects to be an ecology of companies serving an AI economy.
- When I remarked upon how it sure looked to me like humans had an architectural improvement over chimpanzees that counted for a lot, Hanson replied that this seemed to him like a one-time gain from allowing the cultural accumulation of knowledge.

I emphasize how all the mighty human edifice of Go knowledge, the joseki and tactics developed over centuries of play, the experts teaching children from an early age, was entirely discarded by AlphaGo Zero with a subsequent performance improvement. These mighty edifices of human knowledge, as I understand the Hansonian thesis, are supposed to be *the* bulwark against rapid gains in AI capability across multiple domains at once. I was like "Human intelligence is crap and our accumulated skills are crap" and this appears to have been bourne out.

Similarly, single research labs like Deepmind are not supposed to pull far ahead of the general ecology, because adapting AI to any particular domain is supposed to require lots of components developed all over the place by a market ecology that makes those components available to other companies. AlphaGo Zero is much simpler than that. To the extent that nobody else can run out and build AlphaGo Zero, it's either because Google has Tensor Processing Units that aren't generally available, or because Deepmind has a silo of expertise for being able to actually make use of existing ideas like ResNets, or both.

Sheer speed of capability gain should also be highlighted here. Most of my argument for FOOM in the Y-H debate was about self-improvement and what happens when an optimization loop is folded in on itself. Though it wasn't necessary to my argument, the fact that Go play went from "nobody has come close to winning against a professional" to "so strongly superhuman they're not really bothering any more" over two years just because that's what happens when you improve and simplify the

architecture, says *you don't even need self-improvement* to get things that look like FOOM.

Yes, Go is a closed system allowing for self-play. It still took humans centuries to learn how to play it. Perhaps the new Hansonian bulwark against rapid capability gain can be that the environment has lots of empirical bits that are supposed to be very hard to learn, *even in the limit of AI thoughts fast enough to blow past centuries of human-style learning in 3 days*; and that humans have learned these vital bits over centuries of cultural accumulation of knowledge, *even though we know that humans take centuries to do 3 days of AI learning when humans have all the empirical bits they need*; and that AIs cannot absorb this knowledge very quickly using "architecture", *even though humans learn it from each other using architecture*. If so, then let's write down this new world-wrecking assumption (that is, the world ends if the assumption is false) and be on the lookout for further evidence that this assumption might perhaps be wrong.

Tl;dr: As others are already remarking, the situation with AlphaGo Zero looks nothing like the Hansonian hypothesis and a heck of a lot more like the Yudkowskian one.

Added: AlphaGo clearly isn't a general AI. There's obviously stuff humans do that make us much more general than AlphaGo, and AlphaGo obviously doesn't do that. However, if even with the human special sauce we're to expect AGI capabilities to be slow, domain-specific, and requiring feed-in from a big market ecology, then the situation we see without human-equivalent generality special sauce should not look like this.

To put it another way, I put a lot of emphasis in my debate on recursive self-improvement and the remarkable jump in generality across the change from primate intelligence to human intelligence. It doesn't mean we can't get info about speed of capability gains *without* self-improvement. It doesn't mean we can't get info about the importance and generality of algorithms *without* the general intelligence trick. The debate can start to settle for fast capability gains before we even get to what I saw as the good parts; I wouldn't have predicted AlphaGo and lost money betting against the speed of its capability gains, because reality held a more extreme position than I did on the Yudkowsky-Hanson spectrum.

Crossposted here.

Community Capital

Argument: There is a Thing called Community Capital (closely related to social capital), and because it is based more on How Humans Work, you can get a lot more out of it than just plain old Money.

Epistemic Status: Opening a conversation. All these ideas are in sand.

I used to compare the prices of Various Weekend Activities to the pricetag of one of my main hobbies (we'll randomly call it "Ballooning" to keep it simple). I might pay $35 for a weekend of Ballooning, which would include lodging, delicious food and alcoholic beverages, classes and activities, and lots of one-on-one help. Other Various Weekend Activities tend to be... quite a bit more than that. You would think Ballooning events must be operating at a loss, but it turns out that they actually tend to MAKE money. So what's happening here?

My guess is that money is not actually an efficient unit of exchange, as compared to social or community capital. When you buy things with money you are paying people a premium to do tasks they are not inherently motivated to do, often in situations that are not convenient for them. I can do my own dishes in the two minutes that I'm in my kitchen waiting for my tea water to heat up. However if I am paying someone, in most cases they have to schedule a large chunk of time and travel out of their way to my house to do a task they have no motivation to do except for the exchange of money.

People in communities generally have an inherent desire to contribute to their community. They are often already in a position to help. They are preselected for being interested in many of the sorts of tasks the community needs (e.g. a swing dance community needs swing dance instructors and swing dance music djs). They have social ties to the people they are helping, as opposed to being strangers. They gain social status by contributing and thus find it satisfying.

When I go to a Ballooning weekend I am not just spending $35. I may spend 3 hours washing dishes after a meal. I may spend 10 hours prepping material for a class, and one hour teaching it. I may spend 20 hours making items that are going to be given away. I may spend 3 hours at a meeting helping to organize the event. I may spend an hour on setting up or cleaning up. I may spend 5 hours working with a new person one-on-one.

Sure, if I were to value my time at $20/hour, then my $35 Ballooning events would suddenly be a LOT more expensive! But with the exception of dish washing, which I view as "putting in my time", these are all activities I enjoy to some degree, and they feel more like "participating in my fun hobby" rather than "doing work".

I ENJOY making things that will be given to and treasured by my friends. I ENJOY working one-on-one with new people, and watching them fall in love with the things that I love. I ENJOY contributing to the community that means so much to me. If I could instead work an extra hour at my place of employment and then use that money to hire people to pack up... that's not an exchange I'd want to make.

And it's not only time that can be bought more cheaply with community currency than with money. Physical objects can also be bought with the currency of community.

Sometimes you can borrow an expensive piece of equipment rather than buy it. Sometimes I buy nicer equipment for myself, and I enjoy giving away my starter equipment to new people to help get them started. Sometimes items are more easily bought in bulk sizes that you don’t need, and you’d rather give away the excess than store it forever or trash it. Sometimes I want to rent a venue for cheap, and if I have a large community all willing to do some simple searching in their networks, I can find a deal on a space that’s never been posted on a website.

As with all advice, you may want to reverse this. If you have infinite money and very little free time, it makes sense to spend the money rather than the time that is generally required for developing community capital. If you have Very Important Needs that you don't think your community can commit to or handle, it makes sense to buy them with money.

Attributions and Contributions:

- I already had these thoughts but reading a similar comment by ingres inspired me to type them up.
- Thanks to Raymond and Duncan who did read-throughs of the rough draft

Tensions in Truthseeking

Epistemic Effort: I've thought about this for several weeks and discussed with several people who different viewpoints. Still only moderately *confident though.*

So, I notice that people involved with the rationalsphere have three major classes of motivations:

Truthseeking (how to think clearly and understand the world)

Human/Personal (how to improve your life and that of your friends/family)

Impact (how to change/improve the the world at large)

All three motivations can involve rationality. Many people who end up involved care about all three areas to some degree, and have at least some interest in both epistemic and instrumental rationality. And at least within the rationalsphere, the Personal and the Impact motivations are generally rooted in Truth.

But people vary in whether these motivations are terminal, or instrumental. They also have different intuitions about which are most important - or about how to pursue a given goal. This sometimes results in confusion, annoyance, distrust, and exasperated people working at cross purposes.

Terminal vs Instrumental Truth

For some, truthseeking is important because the world is confusing. Whether you're focused on your personal life or on changing the world, there's a lot of ways you might screw up because something seems right but doesn't work or has a lot of negative externalities. It's necessary to do research, to think clearly, and to be constantly on the lookout for new facts that might weigh on your decisions.

For others, truth-seeking seems more like a fundamental part of who they are. Even if it didn't seem necessary, they'd do it anyway because because it just seems like the right thing to do.

I think there's a couple layers of conflict here. The first is that instrumental-truthseekers tend to have an intuition that lots of other things matter as much or more than truth.

> It's more important to be able to launch a startup confidently than to have an accurate perception of its chance of success. It's important not to immediately criticize people because that disincentivizes them from trying new things. Interpersonal relationships seem to need 5x as many compliments as criticisms to flourish. It's important to be able to run a successful marketing campaign, and even if you're trying to run an honest marketing campaign, comprehensive honesty requires effort that might be better spent on actually building your product or something.
>
> It may even be necessary to schmooze with people you don't respect because they have power and you need their help if you're going to make a dent in the universe.

Then, there are people (who tend to be terminal-truthseekers, although not always), who counter:

> Earnest truthseeking is incredibly rare and precious. In almost every movement, it gets sacrificed on the altar of practicality and in-group solidarity. Can't we just once have a movement where truthseeking is the primary thing that never gets sacrificed?
>
> This doesn't just seem worth trying for the novelty of it: the world seems so deeply confusing, the problems it faces seem so immense and so entwined with tribal-politics that distort the truth, that we probably need a movement of impact-oriented truthseekers who never compromise their intellectual integrity no matter what.

I find this argument fairly compelling (at least for a deeper delve into the concept). But what's interesting is that even if it's an overriding concern, *it doesn't really clarify what to do next.*

The Trouble With Truthseeking While Human

On the one hand, social reality is a thing.

Most cultures involve social pressure to cheer for your ingroup's ideas, to refrain from criticizing your authority figures. They often involve social pressure to say "no, that outfit doesn't make you look fat" whether or not that's true. They often involve having overt, stated goals for an organization (lofty and moral sounding) that seem at odds with what the organization ends up doing - and if you try to mention the disconnect between what people are saying and what they are doing, they get upset and angry at you challenging their self-conception.

The pressure to conform to social reality is both powerful and subtle. Even if you're trying to just think clearly, privately for yourself, you may find your eyes, ears and brain conforming to social reality anyway - an instinctive impulse to earnestly believe the things that are in your best interest, so you peers never notice that you are doubting the tribe. I have noticed myself doing this, and it is scary.

In the face of that pressure, many people in the rationality community (and similar groups of contrarians), have come to prize criticism, and willingness to be rude. And beyond that - the ability to see through social reality, to actively distance themselves from it to reduce its power over them (or simply due to aesthetic disgust).

I earnestly believe those are important things to be able to do especially in the context of a truthseeking community. But I see many people's attempts as akin Stage 2 of Sarah Constantin's "Hierarchy of Requests":

> Let's say you're exhausted; you want to excuse yourself from the group and take a nap.
>
> In stage 1, you don't dare ask. Or you don't understand why you feel shitty, you don't recognize it as fatigue. You just get more and more upset until you collapse in a heap. In stage 2, you rudely interrupt people in the middle of something important and announce that you're tired and you're leaving. In stage 3, you find a convenient moment, apologize for cutting things short, but explain that you've got

to get some rest. In stage 4, you manage to subtly wrap things up so you can get some rest, without making anyone feel rushed.

It's better to be able to rudely criticize than not at all. And for some people, a culture of biting, witty criticism is fun and maybe an important end in-and-of-itself. (Or: a culture of being able to talk about things normally considered taboo can be freeing and be valuable both for the insight and for the human need for freedom/agency). I've gotten value out of both of those sorts of cultures.

But if you're unable to challenge social reality *without* brusquely confronting it - or if that is the manner in which you usually do - I think there's a lot of net-truth you're leaving on the table.

There are people who don't feel safe sharing things when they fear brusque criticism. I think Robby Bensinger summarized the issue compactly: "My own experience is that 'sharp culture' makes it more OK to be open about certain things (e.g., anger, disgust, power disparities, disagreements), but less OK to be open about other things (e.g., weakness, pain, fear, loneliness, things that are true but not funny or provocative or badass)."

Brusque confrontation leads to people buckling down to defend their initial positions because they feel under attack. This can mean less truth gets uncovered and shared.

Collaboration vs Criticism For The Sake Of It

The job of the critic is much easier than the job of the builder.

I think that there's a deeper level productive discussion to be had when people have a shared sense that they are *collaboratively building something*, as opposed to a dynamic where "one person posts an idea, and then other people post criticisms that tear it down and hopefully the idea is strong enough to survive." Criticism is an important part of the building process, but I (personally) feel a palpable difference when criticized by someone who shows a clear interest in *making sure that something good happens as a result of the conversation.*

Help Brainstorm Solutions - If you think someone's goals are good but their approach is wrong, you can put some effort into coming with alternate approaches that you think are more likely to work. If you can't think of any ways to make it work (and it seems like it's better to do nothing than to try something that'll make a situation worse), maybe you can at least talk about some other approaches you considered but still feel inadequate.

Active Listening / Ideological Turning Tests - If you disagree with a person's *goals,* you can try to understand why they have those goals, and showcase to them that you at least get where they're coming from. In my experience people are more willing to listen when they feel they're being listened to.

Accompanying criticism with brainstorming and active listening acts as a costly signal, that helps create an atmosphere where it's a) worth putting in the effort to develop new ideas, and b) easier to realize (and admit) that you're wrong.

Truth As Impact

If you constantly water down your truth to make it palatable for the masses, you'll lose the spark that made that truth valuable. There are downsides to being constantly guarded, worried that a misstep could ruin you. Jeff Kaufman writes:

> There are a lot of benefits to unguarded communication: you can move faster, you can open up your tentative thoughts to friendly consideration and criticism, you don't have the mental or process overhead of needing to get every statement as perfect as possible. You might say something that you don't mean to, but in a friendly environment you can correct yourself or accept someone else's correction.
>
> Despite these benefits, it seems to me that things generally move in the more guarded direction, at least publicly, as they become more successful.

Daniel in the comments notes:

> And I think one cost of guardedness that seems missing from the post is that guardedness can bias thinking in favor of more easily palatable and defensible ideas, both in discussions between people as well as one's own thoughts.
>
> Unfortunately, I think it's a natural consequence of growing large and powerful enough to actually affect the big picture: If you're communicating, not to a few trusted friends but to the entire world, then a verbal misstep will *not* be something you can easily correct, and the cost may grow from "a few minutes of clarification" to "millions of dollars worth of value lost".

I'm not sure how to handle that paradox (Less Wrong is hardly the first group of people to note that PR-speak turns dull and lifeless as organizations grow larger and more established - it seems like an unsolved problem).

But there's a difference between watering things down for the masses and speaking guardedly... and learning to communicate in a way that uses other people's language, that starts from their starting point.

If you want your clear insights to *matter anywhere outside a narrow cluster of contrarians*, then at some point you need to figure out how to communicate them so that the rest of the world will listen. Friends who are less contrarian. Customers. Political bodies. The Board of Directors at the company you've taken public.

How to approach this depends on the situation. In some cases, there's a specific bit of information you want people to have, and if you can successfully communicate that bit then you're won. In other cases, the one bit doesn't doing anything in isolation - it only matters if you successfully get people to think clearly about a complex set of ideas.

Consider Reversing All Advice You Hear

One problem writing this is that there's a lot of people here, with different goals, methods and styles of communication. Some of them could probably use advice more like:

- "Learn to criticize more kindly/constructively."
- "Communicate more clearly."
- "Keep in mind the layers of signals you're sending when you try to state 1st-order-true-things."

And some some could probably use advice more like:

- "Make sure in your efforts to avoid conflict you don't gloss over important disagreements."
- "Don't get so wrapped up in what other people think that you lose the ability to think clearly for yourself."

I started writing this post four months ago, as part of the Hufflepuff Sequence. Since then, I've become much less certain about which elements here are most important to emphasize, and what the risks are of communicating half-baked versions of each of those ideas to different sorts of people.

But I do still believe that the end-goal for a "true" truth-oriented conversation will need to bear all these elements in mind, one way or another.

"Focusing," for skeptics.

This is a linkpost for https://medium.com/@ThingMaker/focusing-for-skeptics-6b949ef33a4f

Gendlin's Focusing technique is super rad. I know this because *everybody* keeps telling me so.

(Okay, not quite everybody, but a really tediously large percentage of the people in my online social circle.)

But I've tried it a bunch of times, in a bunch of variants, with a bunch of qualified mentors trying to help, and it's just never clicked. I've listened to the audio book and gone through all the steps, and it just doesn't do anything for me.

So here's my variant—the thing that I do instead, which I claim is using the same hardware and software and providing me with the same kind of improved introspective access. If you're one of those skeptics who thought it all sounded nuts, or one of those unlucky people who thought it sounded *awesome* but could never make it work, this post has your name on it.

The "big idea" of Focusing (according to me) is that parts of your subconscious System 1 are storing up massive amounts of accurate, useful information that your conscious System 2 isn't really able to access. There are things that you're aware of "on some level," data that you perceived but didn't consciously process (see blindsight as both concrete example and metaphor), competing goalsets that you've never explicitly articulated, and so on and so forth.

Focusing is a technique for bringing some of that data up into conscious awareness, where you can roll it around and evaluate it and do something about it. Half of the value comes from just *discovering that the information exists at all* (e.g. noticing feelings that were always there and strong enough to Imperius you but which were somewhat "under the radar" and subtle enough that they'd never actually caught your attention), and the other half comes from having new models to work with and new theories to test. If I manage to recognize that e.g. a significant chunk of my romantic problems stem from self-censorship pressure because of a strong aversion to seeming needy, I suddenly have threads to pull on rather than falling back to just "Yeah, things with Cameron are not great."

The actionable claim of Focusing (again, according to me) is that this information expresses itself in "felt senses" in the body—think butterflies in the stomach, or your throat closing up, or the heat of embarrassment in your cheeks, or a heavy sense of doom that makes your arms feel leaden and numb, or whatever physiological sensation happens to you when you catch yourself about to tell a lie. The brain doesn't know how to drop its information directly into your verbal loop, so instead it falls back on influencing your physiology and hoping that you notice (or simply respond).

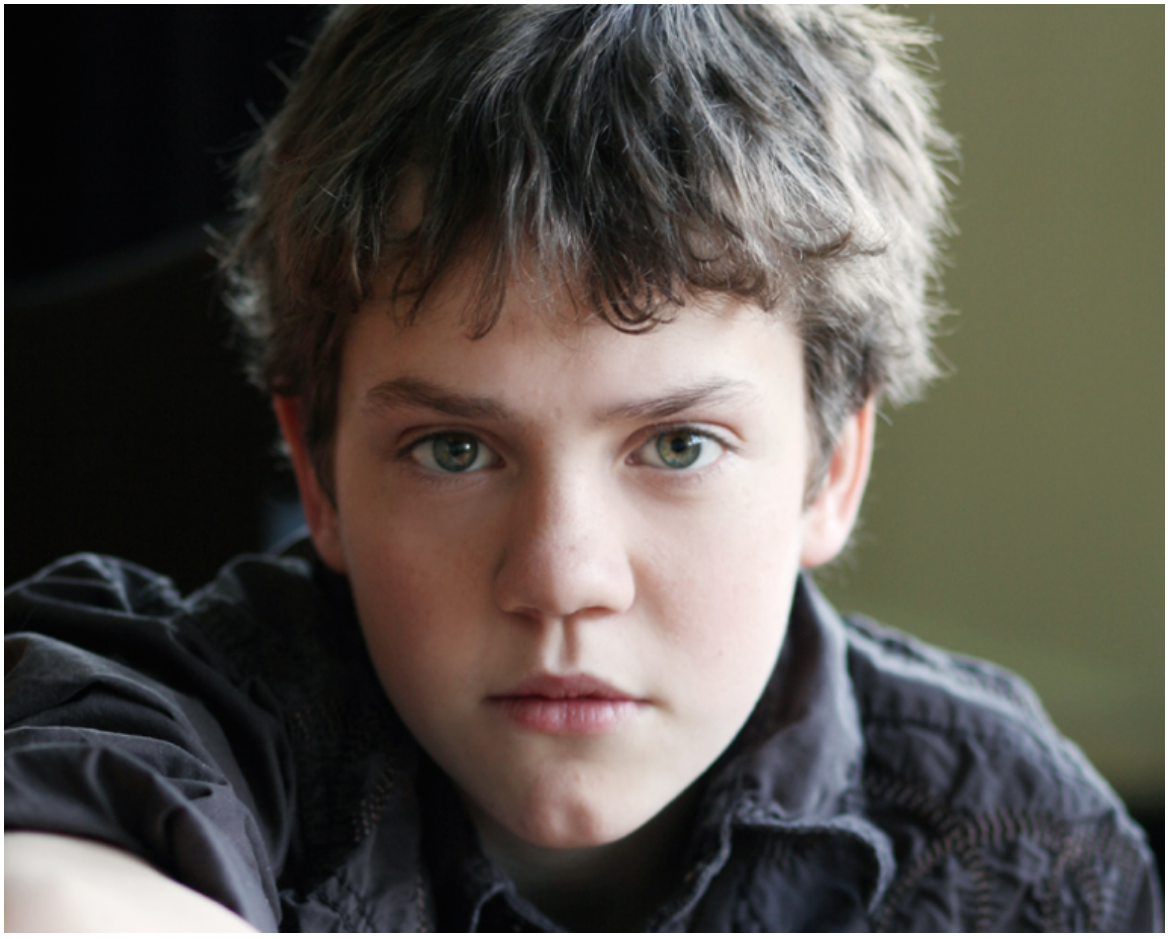
Gendlin recommends a series of steps that help you build up the skill of noticing and dialoguing with these felt senses until they yield their precious data (often changing or disappearing in the process).

I'm going to focus (tee hee) on the one part of Gendlin's process that makes the most sense to me, which is *finding the felt sense's True Name.* In the official algorithm, this is step three—finding a handle. In my version, it's basically the whole technique.

Note: it's worth mentioning that when Gendlin titled his process "focusing," he meant it in the sense of "gently turning the knob on a microscope to bring things into focus," and

definitely not *in the sense of* "*buckle down and try real hard to effortfully bring your attention to bear.*"

All right, diving in. First, take a look at this face.

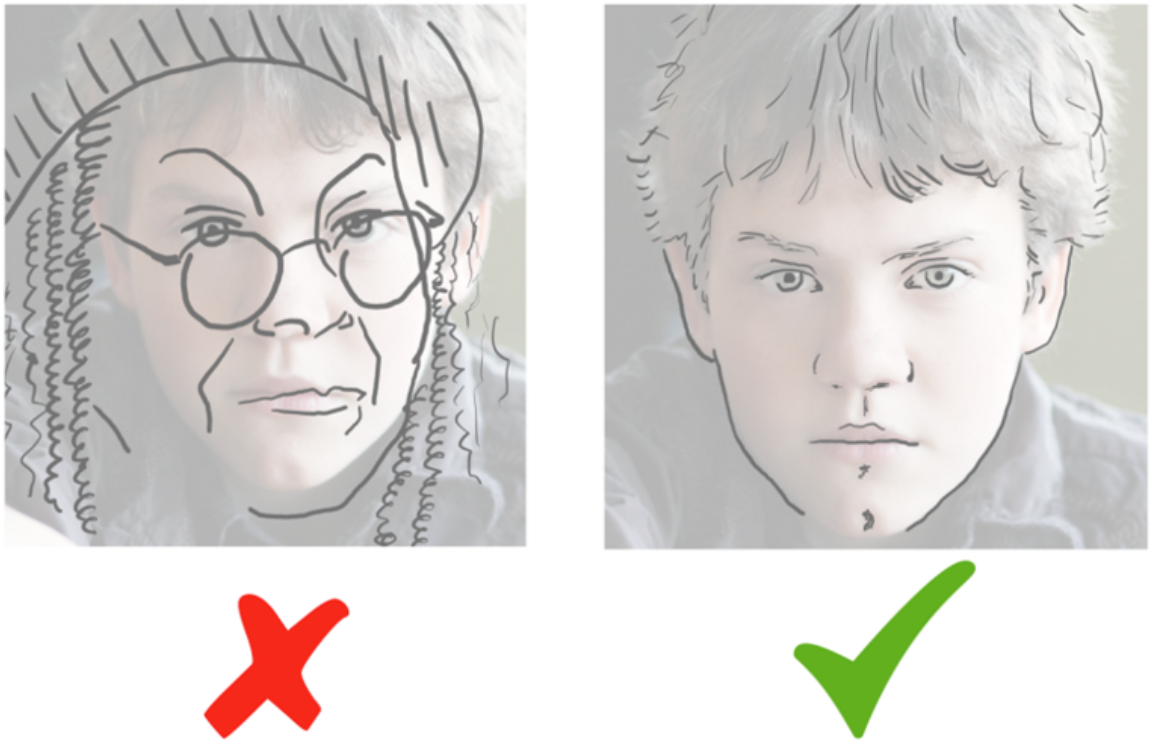


A particular face making a particular expression is going to be our metaphor for a felt sense. Faces and expressions are rich in contrast and detail, they're extremely specific and recognizable, and they're very hard to describe in words—just like my implicit models of the dynamics between me and Cameron and all of our history and all of my unstated assumptions about how romantic relationships work.

(Side note: Cameron isn't real. Cameron is like Maria in a Counting Crows song.)

A *sketch,* on the other hand, is *compressed.* It can be evocative, but it's sparse and utilitarian, conveying as much of the relevant information as possible with economy of line. In order to get something as rich as a real face out of a sketch, your brain has to do a lot of processing, and regenerate a lot of information from cached models and past experiences.

What’s most important for our metaphor is the concept of trueness. Fit, accuracy, veracity—the quality of an actual correspondence between model and reality. It’s not about how detailed or technically sophisticated the sketch is, it’s about whether it *matches what it’s trying to match*.



The relationship between [words] and [felt senses] is analogous to the relationship between [sketches] and [faces]. You could riff off of the old saying and claim that “a felt sense is worth a thousand words.”

And the practice of Focusing (or at least my version of it) is one of using careful introspective attention to zero in on the *right* words. The name or “handle” you come up with won’t be anywhere near the whole story, but it can nevertheless be the short, compressed version of the *right* story.

(The same principle applies to “partial” sketches—you could imagine just a few lines around the nose and one side of the jaw that *don’t* convey the whole picture, but *do* very accurately match a part of it. In our metaphor, that would be like gaining clarity on one aspect of the thing that’s bothering you, even if you still can’t see the whole picture.)

(Also, while the rest of this post focuses on words specifically, it's worth noting that you could engage in a similar process to translate your felt senses into any other medium, as well. e.g. a vivid mental image, or a scribbled picture, or a free-verse poem, etc. If words aren't working for you, try a different modality.)

Okay, let’s try it. We’ll bounce back and forth between words and pictures to try to give you a sense of what mental motion I’m making.

> *Okay, so there’s clearly SOMETHING bothering me. And it’s got something to do with Cameron.*

This is like knowing “the picture is a face” as opposed to maybe “the picture is a fighter jet.” It's a very coarse starting point, trying to get us in the right ballpark.

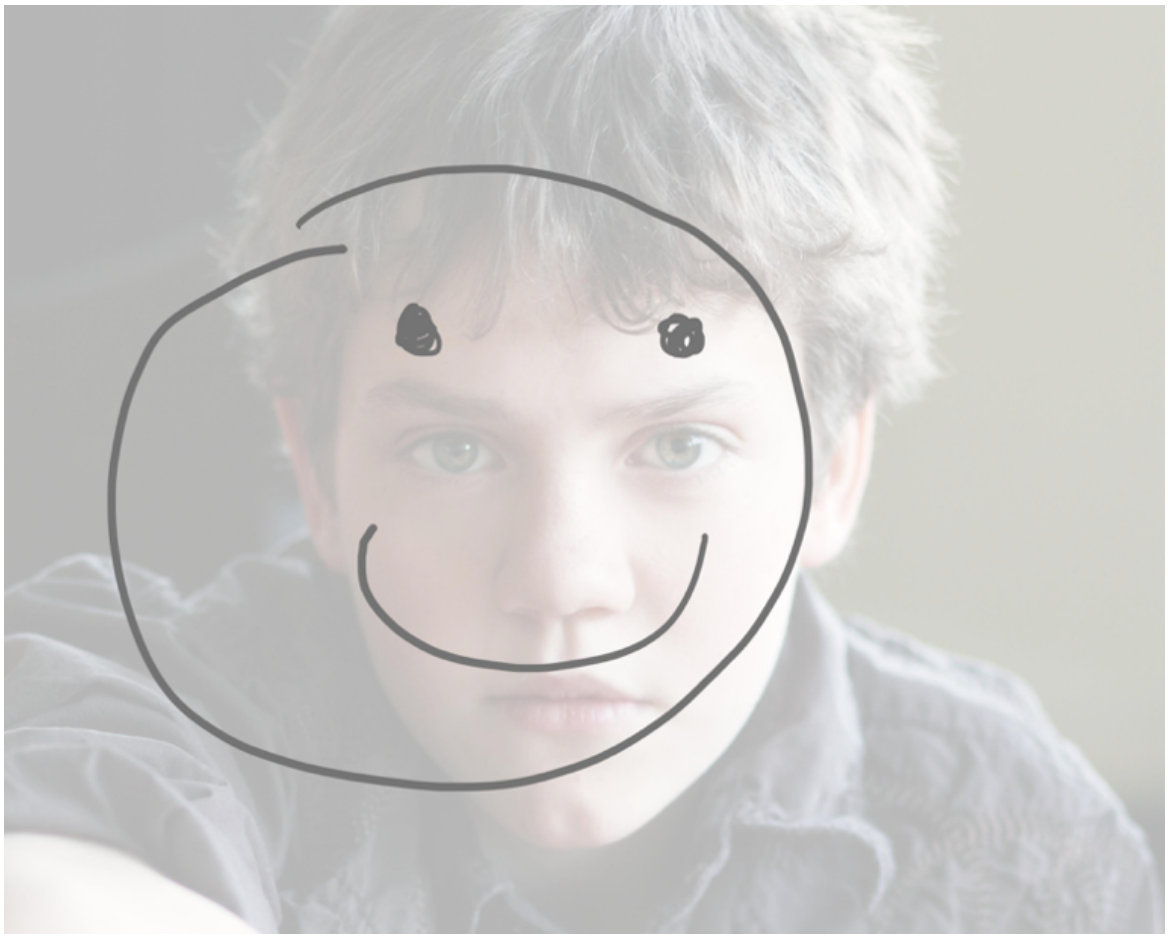
At this point, it helps to get a clear sense of *where the problem lives in my body,* i.e. does this feeling express itself in any identifiable physical sense? What happens to my subjective physiological experience, when I turn my attention to "the thing that's bothering me about my relationship with Cameron"? Does the reaction live in my belly, in my limbs, in my head? Do I experience it as an ache, a tingle, a heat, a chill?

To be clear, you're not actually trying to pin the sensation down and put words on it, at this point—you're just making sure that you know *which* feeling you're dealing with. If you have multiple stressors in your life, they will likely show up in different ways, and it helps to be dialoguing with a single felt sense at a time.

It's also worth noting that a lot of people struggle with Focusing *specifically because* they have a hard time zeroing in on a physiological marker, and according to me, *that doesn't mean you can't do Focusing.* I think you can do something analogous with e.g. your sense-of-what's-true—you can make an explicit claim about the problem, and then check whether that claim seems accurate, without ever referencing sinking feelings in your stomach (or whatever).

So, first iteration:

> *Have we been fighting a lot?*



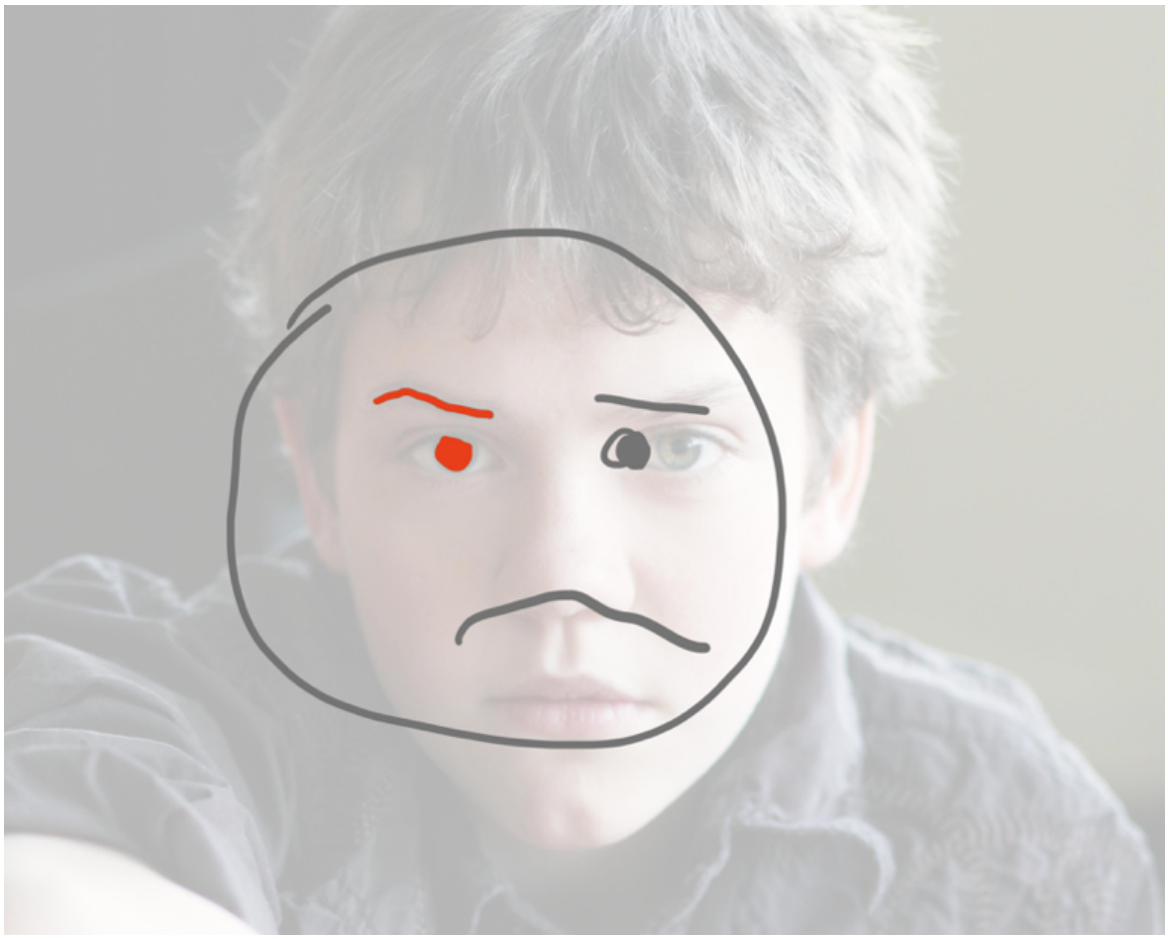
No, that's not it at all.

(This is the equivalent of saying "no, not a *smiley* face, a *serious* face.")

I got that reaction by *holding up* the hypothesis/potential handle against the feeling and comparing them.

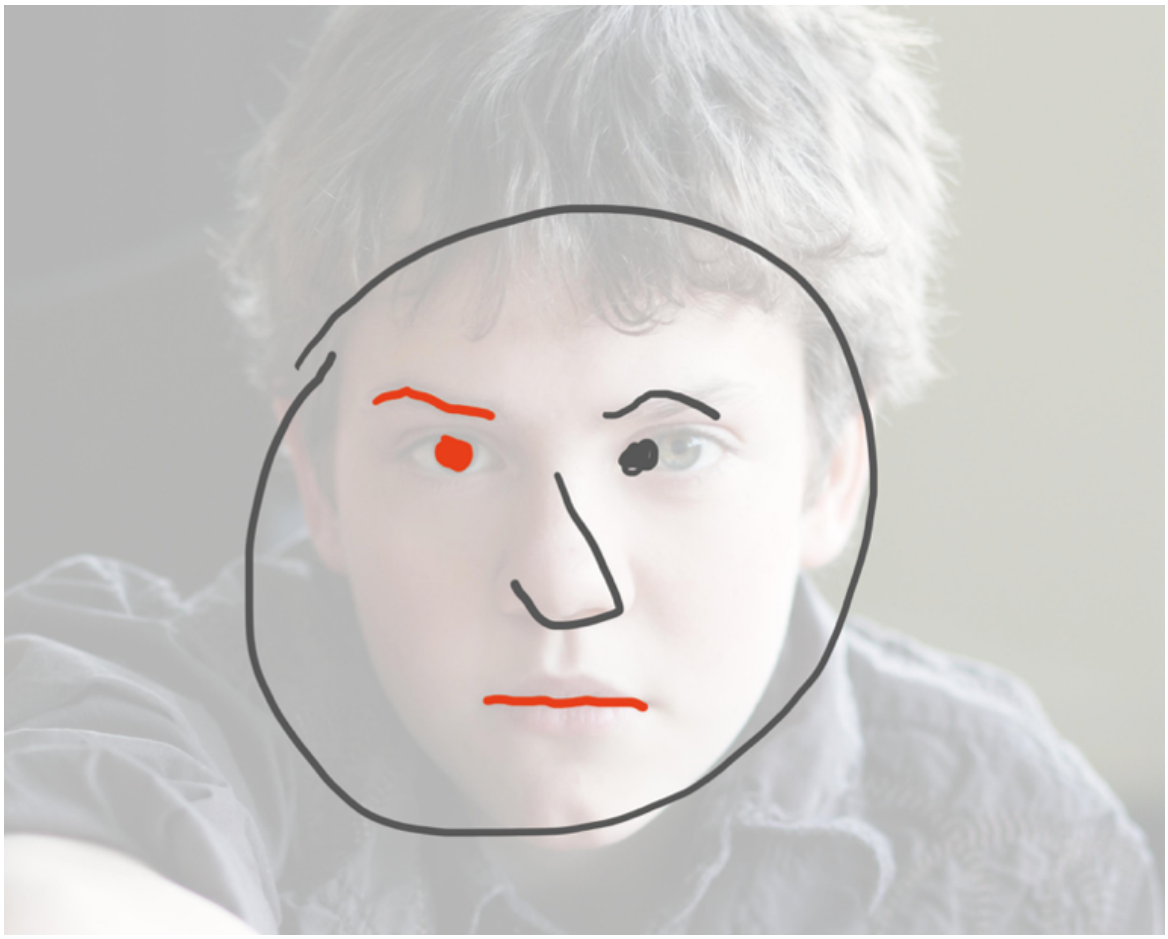
It's more like — like — ugh, like I never know what to say?

*No, it’s like I **have** to say the right things, or else.*

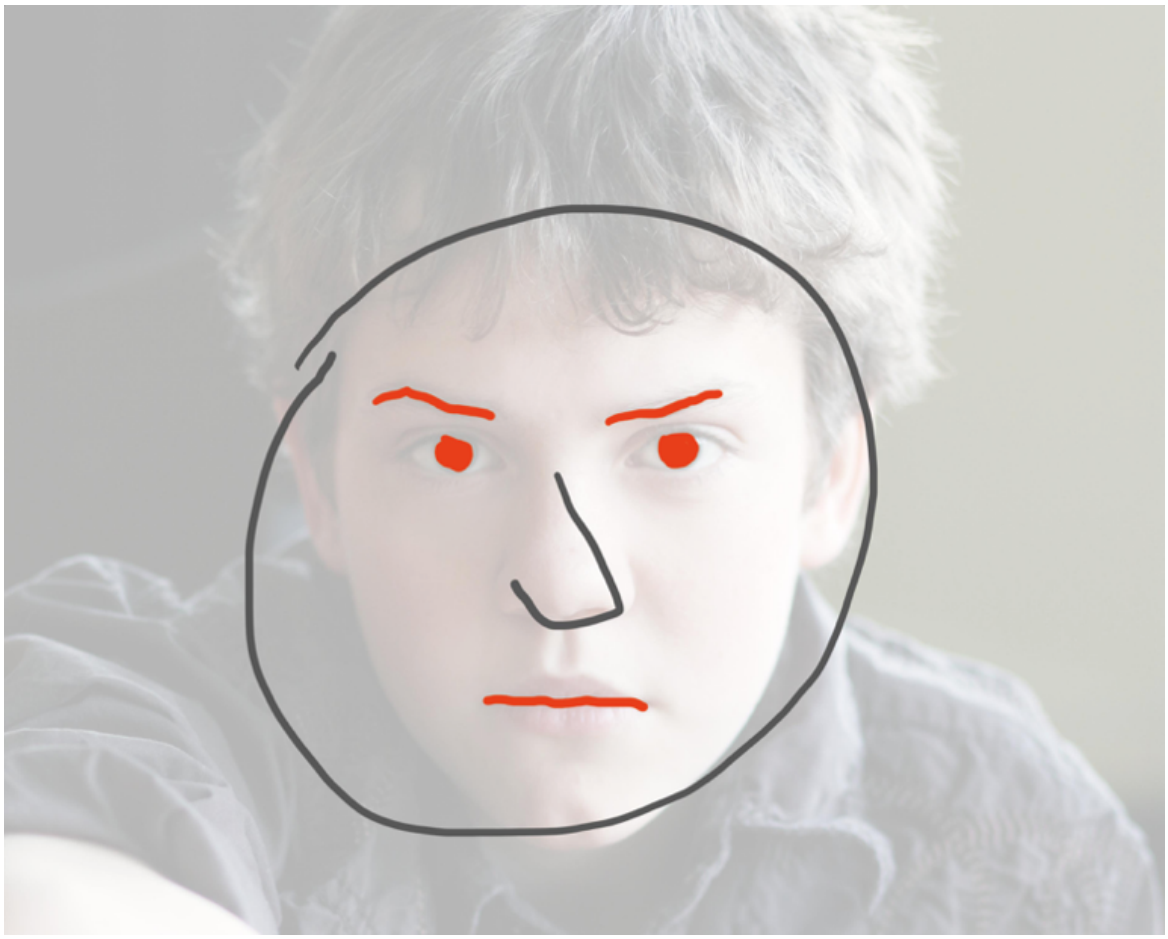


(Here I've gotten a *partial* match—I'm nowhere near the real story, but *something* about that handle I just tried out resonated. It has the right flavor, and I can sort of poke around *near* the thing I just said for other words that might be a closer fit. If the initial flailing about was the "stochastic" part, now I'm in "gradient descent." Something has responded, something has changed—the nonverbal part of my brain that's trying to send up a message is telling me "yes, you're getting warmer!")

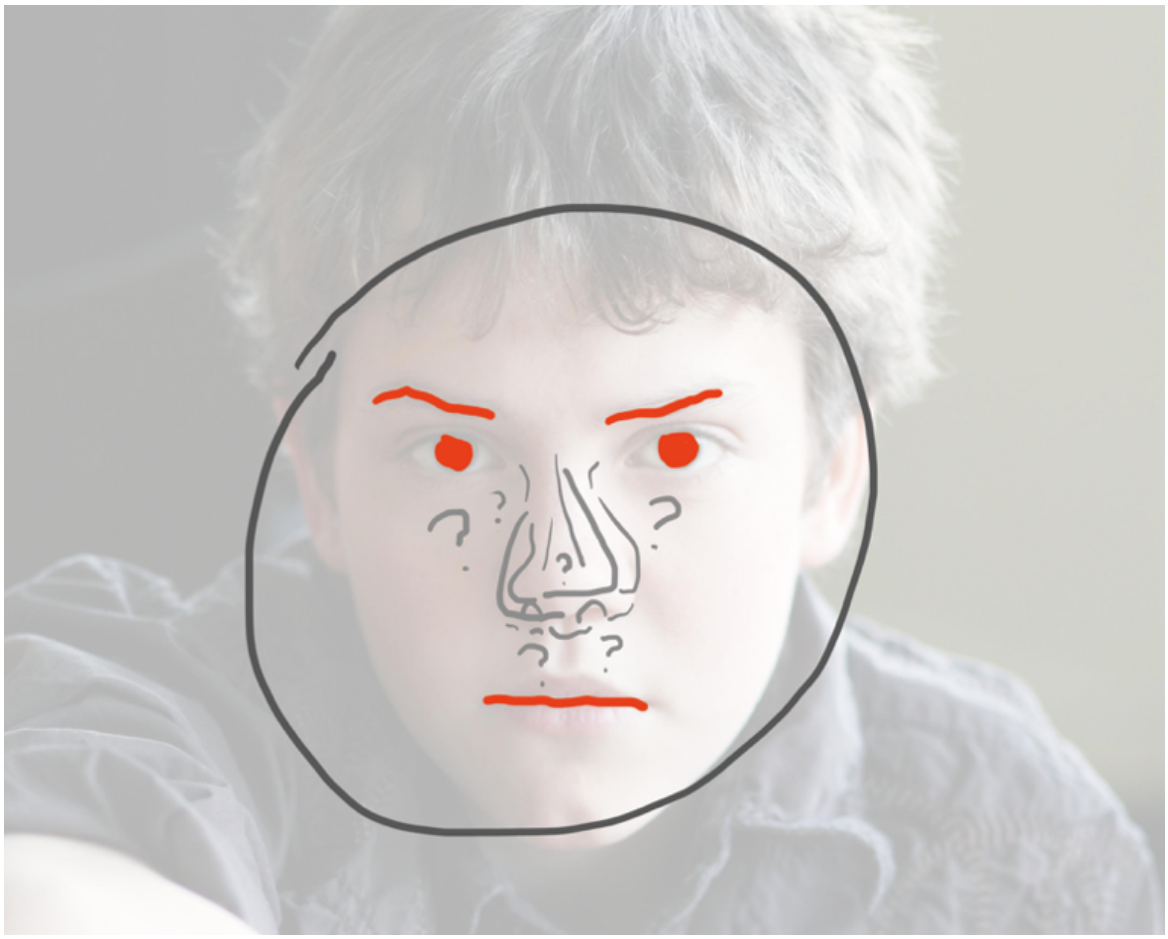
*It's like — if I say the **wrong** thing, then everything falls apart and it's all ruined.*



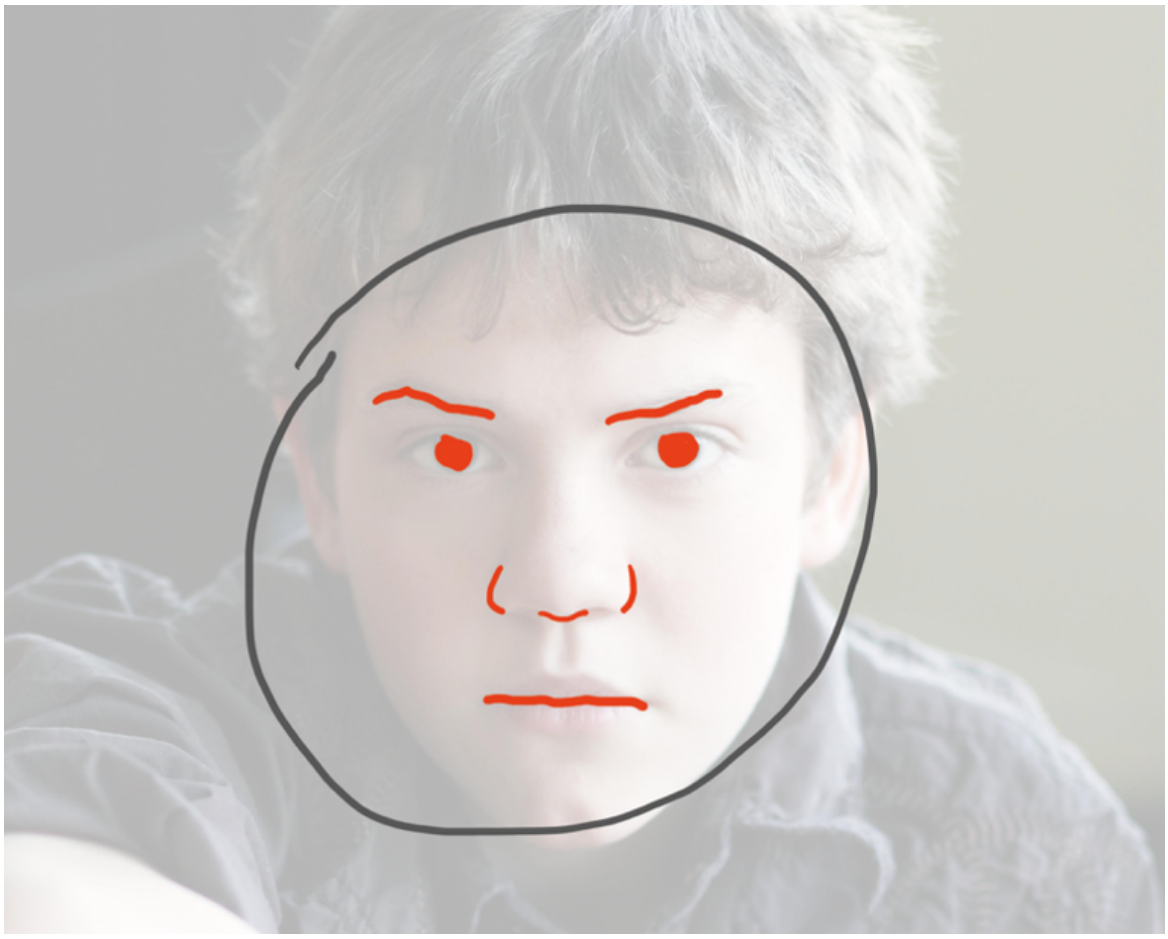
*And it’s like I’m the only one? Like, **Cameron** doesn’t have to pay attention, just **me**. Cameron gets to —*



— to —



*— to **relax.** That’s it. Yeah. It feels like I’m the only one who doesn’t get to relax.*



(There, I was doing a super-fast scan over a whole bunch of possible words and phrases and explanations, all the while paying very close attention to my truth-detection module. I knew where the answer *would* be, and I knew its general shape, but I had to keep looking until I found something *juuuuust* right. It's like when you mentally stutter past five or six different comebacks to throw at your classmate until you find the one that's cutting, true, *and* okay-to-say-even-though-your-teacher-is-listening.)

You get the idea. As the process continues, the verbal sketch grows more and more accurate, and *evokes* more and more of the underlying what's-really-going-on. I can feel a sort of click, or a release of pressure, or a deep rightness, once I say the thing that *really* completes the picture.

(In Gendlin's Focusing, he makes the point that the felt sense will often change—or vanish—once you've given it the right handle. It's sort of as if the part of you that was trying to send up a red flag and expressing itself in a physiological sensation *no longer has to keep sending that signal* once your conscious brain receives the message. There's often a relaxing, opening-up sort of feeling once the unrecognized problem becomes explicitly recognized.)

> *The problem is, I feel like I'm the one who has to be the grownup, because if Cameron gets mad and I **don't** fix it, then that's it—things are just bad forever. But when **I'm** the one who gets mad, Cameron doesn't do jack squat. Cameron protects Cameron, and I protect **us**, and that's something I never get to put down. It feels like I'm the only one who's putting effort into maintenance, and that makes me feel like Cameron doesn't really want the relationship at all, or at least doesn't **value** it and wouldn't miss it if it*

*ended, and that's lonely and exhausting and it makes me feel like maybe **I'm** not worth valuing.*

I've glossed over a few things in describing this process, to save time. For instance:

- As I mentioned briefly above, the handle doesn't always come in words—sometimes it can be a vivid image, or a metaphor, or a poem, rather than a straightforward description.
- Often, there'll be a whole other felt sense (or three) that arises and needs to be either a) dealt with or b) set aside
- Frequently, a given attempt to iterate on finding a handle—a given single step in the process above—will move your overall understanding *away* from accuracy rather than toward it. That doesn't mean you're doing it wrong. Patience is key.

And of course, just because you've accurately named *your brain's sense of what's going on* doesn't mean that you've found the actual truth—it's almost trivially easy to construct another side to this story in which Cameron's trying to relate to me as an equal, but I insist on being patronizing and untrusting and laying blame for things Cameron *never asked me to do, why does it always have to be this huge deal, why can't we just have fights every now and then like normal people and just **get over it** without acting like it **means** something? Now it's not only that you've pissed me off, it's also that I'm not **allowed** to be angry—that I **have** to calm down and make up or you'll start acting like our whole relationship is doomed. That's some serious hostage-taking right there.*

Or maybe not. Maybe that isn't it, and the real problem is on a completely different axis. But either way, I've gotten clarity on what was going on in *my* head, under the hood—on what sorts of narratives and frames resonate with the part of my subconscious that was generating the feelings of frustration or unease in the first place. That's a *huge* step forward in turning the problem into something tractable, with gears and levers and switches. Instead of being a Mysterious Mystery, it's now mundane (which is not to say that it'll be *easy to fix*, just that I'm no longer fumbling around in the dark).

To me, that's super valuable, even if a lot of the people I know who do Focusing insist I'm doing it wrong.

Social Choice Ethics in Artificial Intelligence (paper challenging CEV-like approaches to choosing an AI's values)

This is a linkpost for https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046725

Placing Yourself as an Instance of a Class

There's an intuition that I have, which I think informs my opinion on subjective probabilities, game theory, and many related matters: that part of what separates a foolish decision from a wise one is whether you treat it as an isolated instance or as one of a class of similar decisions.

A simple case: someone doing something for the first time (first date, first job interview, etc) vs someone who has done it many times and "knows how these things go". Surprise events to the greenhorn are tired stereotypes for the old hand. But, sometimes, we can short-circuit this process and react wisely to a situation without long experience and hard knocks.

For example, if a person is trying to save money but sees a doodad they'd like to buy, the fool reason as follows: "It's just this one purchase. The amount of money isn't very consequential to my overall budget. I can just save a little more in other ways and I'll meet my target." The wise person reasons as follows: "If I make this purchase now, I will similarly allow myself to make exceptions to my money-saving rule later, until the exception becomes the rule and I spend all my money. So, even though the amount of money here isn't so large, I prefer to follow a general policy of saving, which implies saving in this particular case." A *very* wise person may reason a bit more cleverly: "I can make impulse purchases if they pass a high bar, such that I actually only let a few dollars of unplanned spending past the bar every week on average. How rare is it that a purchase opportunity costing this much is at least this appealing?" ***does a quick check and usually doesn't buy the thing, but sometimes does, when it is worth it***

One way to get this kind of wisdom is by spending money for a long time, and calibrating willingness-to-spend based on what seems to be happening with your bank account. This doesn't work very well for a lot of people, because the reinforcement happens too slowly to be habit-forming. A different way is to notice the logic of the situation, and think *as though* you had lived it.

Similarly, although people are heavily biased to treat vivid examples from the news (airplane crashes involving celebrity deaths) or personal anecdotes from friends and family (an uncle who fell ill when a high-voltage power line was installed near his home) or worse, from strangers on the internet (the guy who got "popcorn lung" from smoking an e-cig), actually, statistics are far more powerful. (It's true that medical anecdotes from family might be more relevant than statistics due to genetic factors shared with family members, but even so, taking the "statistical view" -- looking at the statistics available, including information about genetic conditions and heritability if available, and then making reasonable guesses about yourself based on your family -- will be better than just viewing your situation in isolation.)

I won't lecture much on the implications of that -- most readers will be familiar with the availability heuristic, the base-rate fallacy, and scope insensitivity. Here, I just want to point out that putting more credence in numbers than vivid examples is an instance of the pattern I'm pointing at: placing your decision as an instance of a class, rather than seeing it in isolation.

In an abstract sense, the statistical view of any given decision -- the view of the decision as part of a class of relevantly similar decisions -- is "more real" than an attempt to view it in isolation as a unique moment. Not because it's actually more real, but because this view is closer to your decision. Humans may exist in a single, deterministic universe -- but we *decide* in statistical space, where outcomes come in percentages.

When the weatherman says "70% chance of rain" in your area, but it is *already* raining outside, you know you're one of the 70% -- he's speaking to an area, and giving the percent of viewers in the area who will experience rain. He can't just say 100% for those viewers who will experience rain and 0% for those who won't. Similarly, when you make a decision, you're doing it in a general area -- you can't just decide to drive when you won't crash and avoid driving when you will (though you can split up your decision by some relevant factors and avoid driving when risk is high).

This exact same intuition, of course, supports timeless/updateless decision theory even more directly than it supports probabilism. Perhaps some future generation will regard the idea of timeless/updateless decision-making as more fundamental and obvious than the doctrine of probabilism, and wonder why subjective probability was invented first.

Trope Dodging

Epistemic status: 2 1/2 hours of hammering out an idea that's been in my head for a while. Not too novel, yet a few useful points at the end. Rehashes some ideas from A Human's Guide to Words.

I think that it is useful to model people as having filters that exist somewhere in their decision making process, filters which ask the question, "Does what I'm about to do match a behaviour trope that I've blacklisted?"

If the answer is yes, then the potential action gets vetoed and the brain goes back to the drawing board. If the answer is no, then the potential action gets bumped further up the decision making chain.

The effect of some filters are more visible than others. What many people would call one's "social filter" is something that normally only vetoes doing or saying something after the potential action has already been brought to your conscious attention. You think about the accurate remark you could make about your colleague's outfit, but then don't say it because your filter deemed it rude.

Ugh fields look a lot like filters that get stronger and stronger with time and operate in the no-man's-land between your opaque subconscious processes and your visible conscious ones.

If you want to be more specific about what it means to "match a trope", we can replace trope with similarity cluster, and say that a potential action matches the trope if it is within a certain distance of the epicenter of the similarity cluster that a trope references.

For my personal use, I've called this sort of filtering Trope Dodging, but if another name seems better, I'm happy to concede.

Here's what this could look like in day to day life:

Sasha thinks that whining and complaining are just terrible ideas. She grew up with siblings that always complained and who never tried to make their lives better. This left a strong impression on Sasha, and now she has a filter that keeps her from doing things that are too similar to the trope of "being whiny". She doesn't even like to discuss with her classmates about the difficulty of the various university course they are enrolled in, because her filter registers that as too close to "being whiny".

You, being a lovable well read rationalist, can probably guess how these sorts of filters are less than optimal. Even if if you have determined that it really is in your best interest to not do anything that would be a prototypical example of a blacklisted trope, reality is messy enough that you are always going to be filtering out actions that would actually be a great idea, but just happen to be within the scope of your filter.

Hmmmm, actually, if filters you to cut out big swathes of "definitely bad ideas", might the time saved in reaching conclusions faster outweigh the loss of superior alternatives?

That’s a totally reasonable hypothesis. If you had a filter that only started to get the wrong answer at the very edges of it's blacklisted trope. You get problems when you have filters that are grossly miscalibrated and filter out a non-trivial amount of useful options.

When you have filters with that broad of a scope, it’s easy to catch something which vaguely matches the aesthetic of a trope, while still being devoid of the “core essence” that caused you to define the trope in the first place. Discussing the difficulty of your classes sort of fits the general aura of “whining and complaining”, but lacks the essential traits of whininess, i.e. a sense of entitlement and a defeatist attitude. There’s nothing inherently wrong with making someone aware that you have wronged them, yet it sort of vaguely smells like “getting angry at people”. I know a lot of people who don’t take time for things like sleep, exercise, and nutrition, because they have such a broad filter against “wasting time”.

In all of these scenarios someone loses out on doing something that could genuinely help them, all because their filters are too broad.

***

So far, you might not be very impressed. Most of what I’ve said could be summed up with, “When people use generalizations to guide their actions, they will often make the wrong decision.” However, I think there are some specific claims I’m now able to make which would have been much harder to make clear without the build up.

Claim 1: By default, one’s filters are likely to be poorly calibrated, and the scope of a filter is proportional to how strongly you feel about the prototypical example of the filter’s trope:

Obviously, people are capable of nuanced views and well calibrated filters, but I think it would be an error to assume that any filters you haven’t given any particular attention to are going to be quality. The initial growth of a filter looks a lot like someone over-steering away from something they decided they didn’t like. If you really don’t like the trope you are trying to avoid, I think that your filter defaults to having an equally large scope.

Claim 2: A poorly calibrated filter does more damage the deeper into your decision making process it lives.

Basically, if a poorly calibrated filter is acting within you range of conscious awareness, then you have decent chance at realizing that something is going on. Being privy to the process of filtering makes it easy to question the quality of the filter.

When you have a filter that acts outside of your conscious awareness, it feels like you just never even think of the ideas that are being filtered. For me, it is very rare for the thought, “Hey, maybe I should ask someone for help” to come to mind, even when it objectively makes sense given the situation. I’ve learned to mostly follow through when someone brings up the idea of getting help, but it’s still something I’m subconsciously filtering against, and it’s hard for the idea to even come to mind.

***

So… given all that, how should one proceed? I don’t have any particular advice on how to calibrate a filter you’ve already identified as a problem. However, combining claim one and two gives some insight into how to seek out these filters. To find poorly

calibrated filters that don’t operate inside your conscious awareness, focus on the things you feel most strongly about. What are things that are very similar to things that you hate, but that aren’t actually harmful, and what are things that are similar to things you really love, but that are actually harmful?

Prompt for discussion: Once one finds deeply rooted filters with poor calibration, how should you go about fixing them?

De-Centering Bias

Summary: A perspective that synthesises biases with other considerations incl. game theory, virtue ethics and knowledge of your limitations.

Intro: Adopting a way of thinking in which you are aware of your own biases is clearly important and one of the areas of rationality that is most based solid evidential grounds. However, this also needs to be reconciled with evolutionary psychology arguments that are psychological functioning has evolved to maximise our reproductive fitness, a big part of which is survival. This post attempts to synthesise these two views. I would also like to suggest a new term, De-Centering Bias to describe this technique given how it marks a shift from bias being the central explanation, to bias having to share the stage with other considerations. The best way to understand this is by example:

(This post was originally title Post-Bias Thinking (or Post-Bias-ism). Unsurprisingly, this was controversial, so I decided to rename it.)

Example 1, one of the most famous experiments in psychology is the Stanford Marshmallow Experiment, which was originally interpreted as showing that children who could resist eating one marshmallow now for the promise of two later tended to have better life outcomes. However, later interpretations showed that this could actually be rational for children in some environments where such promises were not reliable and that the presence of such an environment could explain these results by being a common cause of both effects.

Example 2, often revenge will make people take actions that harm both them and the other person. If we model an actor as a self-interested rational agent then we will come to the conclusion that they are being "irrational". On the other hand, if an actor is willing to go to such extreme lengths to punish someone who wrongs them, then there is a strong incentive not to wrong them in the first place (Scott has argued that it could be considered charitable because it also created a disincentive within the wider community). In the Most Convenient World, the actor will gain the benefits of such a threat existing, whilst never having to carry out their threat.

Example 3, the sunk cost fallacy is the tendency of humans to want to continue a project that they have invested a lot of resources (time, money, effort) into, even if the project is not valuable enough to be worth the resources required to finish the project. When discussing this fallacy we need to be aware that human are not rational agents in that we will often be too lazy to engage in activities that would be worthwhile. Wanting to continue projects in which we have invested large amounts of time in allows us to counter this tendency. So if we were able to press a button and remove sunk cost considerations from our brain, I would not be surprised if this was to make us less effective as an agent (as Elizier says, it is dangerous to be half a rationalist, link, there's a better link somewhere, but I can't find it). But further than that, taking a Virtue Ethics approach, every time you complete a project, you become more like the kind of person who completes projects, so sometimes it might be worth completing a project just so that you completed it, rather than for the actual value the project provides. In this case, this bias seems to make us more rational by mitigating a different way in which we are irrational.

De-centering Bias is not:

The belief that "Bias-Centered Thinking" is wrong all or even most of the time, as opposed to being a challenge that forces us to refine our thoughts further.

The belief that *all* biases have benefits attached. Sometimes attributes evolve merely as side effects. Sometimes they harm us, but not enough to affect our reproductive success (h/t Alwhite).

Limited to game theoretic considerations. See the discussion of the sunk cost fallacy above.

In conclusion, De-Centering Thinking is incredibly simple. All I'm asking you to do is to stop and pause for a second after you've been told that something is a bias and think about whether there are any countervailing considerations. I believe that this is an important area to examine as you could probably fill an entire sequence by expanding this analysis to different biases.

Suggestions for Further Discussion: What is something that is generally considered a bias or fallacy, but where you believe that there are also other considerations?

The Strengths of the Two Systems of Cognition

The two systems of cognition a la Kahneman each seem to each have roles that they fulfill better than the other and in which, for optimal performance, the other system ought not to interfere. Until recently, I had only really thought about when System 2 ought to override System 1, but I believe other cases are worth considering as well.

It is important to be able to get System 2 to shut up and let System 1 work in certain circumstances. Physical movement is an excellent example of this. When you shoot a basketball, throw a football, or use a video game controller you may notice that when you are “on fire” you are not focusing at all on the mechanics of what you are doing. In fact, if you start thinking about what you are doing better than usual and begin analyzing the mechanics of the movements, you will likely begin performing worse. Similarly, when you focus on some specific motion that is supposed to contribute to a correct movement, applying the conscious effort of System 2 to force yourself to do it, you will generally exhibit worse performance than if you just hold an image in your mind of what the movement should generally look like.

The override of System 2 causes you to neglect other important factors that go into a movement: and there are too many for system 2 to think about at once. The strength of System 1 is that it can handle all those factors without trouble.

Thus, we can see where it is important to learn the skill of not trying too hard or focusing conscious attention too forcefully. We often cannot consciously keep in mind all the necessary factors, though we are capable subconsciously, as long as System 2 does not interfere with the process by introducing an override in one specific area. While learning this skill, it is important that one not try too hard to not try too hard. Thus, it can be a very difficult skill to learn, I think particularly for analytical people. See The Inner Game of Tennis for a more thorough exposition of this idea.

The skill of effortless effort, is a very useful one, though, I think, not quite as important as learning how to get your mind to actually try, with System 2 reevaluating the lazy heuristics of System 1, which is one of the necessary skills for long term rational decision making. System 2 overriding System 1 is necessary for dealing with many cognitive biases. To avoid allowing representativeness to alter my judgment of what someone does for a living based on a description of them, I must avoid the impulse to merely match the description with a relevant stereotype and output an answer. System 2 must recognize the need to apply base rates and Bayes Theorem, and remember that base rates are probably better evidence of what someone does than the stereotypes that mesh best with a possibly unreliable description. LessWrong is full of other examples of when this skill is important and why.

Then there is the skill of using both systems in unison, each doing the job it can do well and not interfering with the job of the other system and worsening overall performance.

Take a martial artist as he fights his opponent. He must allow his System 1 to handle the mechanics of how his body moves—the complexities that he has trained for throwing a perfect punch are far more numerous than he consciously is aware. If he focuses too hard on one particular aspect of the strike, such as proper shoulder

rotation in his punch, he may lose fluidity, speed, or power. System 2 must not focus on how he performs each strike. At the same time System 2 is very active, evaluating his opponent’s movements, her favored strikes, how she responds to feints. System 2 creates plans for exploiting any perceived weakness. And in this area System 1 must not enter, System 1 sees a fight and becomes angry, abandons caution and begins using too much power, a mode of fighting that will be quickly punished by an experienced opponent.

When the two are in balance, the martial artist will experience a state of transcendence, other concerns falling away as all attention is focused on the challenge in front of him. I have found this sort of experience tends to provide the greatest fun density per time as compared to anything else I do.

I am not sure if this is the most important of the skills, I think getting System 2 to rethink System 1’s overly hasty judgments is still likely more important if one has goals other and higher than personal enjoyment. But getting both systems working together in unison feels the best of any of the skills and I think engaging in activities that allow for both to be used may be an excellent way of increasing the amount of fun you have in the limited time you have to spend.

Different Worlds

I.

A few years ago I had lunch with another psychiatrist-in-training and realized we had totally different experiences with psychotherapy.

We both got the same types of cases. We were both practicing the same kinds of therapy. We were both in the same training program, studying under the same teachers. But our experiences were totally different. In particular, all her patients had dramatic emotional meltdowns, and all my patients gave calm and considered analyses of their problems, as if they were lecturing on a particularly boring episode from 19th-century Norwegian history.

I'm not bragging here. I *wish* I could get my patients to have dramatic emotional meltdowns. As per the textbooks, there should be a climactic moment where the patient identifies me with their father, then screams at me that I ruined their childhood, then breaks down crying and realizes that she loved her father all along, then ???, and then their depression is cured. I never got that. I tried, I even dropped some hints, like "Maybe this reminds you of your father?" or "Maybe you feel like screaming at me right now?", but they never took the bait. So I figured the textbooks were misleading, or that this was some kind of super-advanced technique, or that this was among the approximately 100% of things that Freud just pulled out of his ass.

And then I had lunch with my friend, and she was like "It's so stressful when all of your patients identify you with their parents and break down crying, isn't it? Don't you wish you could just go *one day* without that happening?"

And later, my supervisor was reviewing one of my therapy sessions, and I was surprised to hear him comment that I "seemed uncomfortable with dramatic expressions of emotion". I mean, I *am* uncomfortable with dramatic expressions of emotion. I was just surprised he noticed it. As a therapist, I'm supposed to be quiet and encouraging and not show discomfort at anything, and I was trying to do that, and I'd thought I was succeeding. But apparently I was unconsciously projecting some kind of "I don't like strong emotions, you'd better avoid those" field, and my patients were unconsciously complying.

I wish I could say my supervisor's guidance fixed the problem and I learned to encourage emotional openness just as well as my colleague. But any improvement I made was incremental at best. My colleague is a bubbly extravert who gets very excited about everything; I worry that to match her results, I would have to somehow copy her entire personality.

But all was not lost. I found myself doing well with overly emotional patients, the sort who had *too many* dramatic meltdowns to do therapy with anybody else. With me, they tended to give calm and considered analyses of their problems, as if they were lecturing on a particularly boring episode from 19th-century Norwegian history. Everyone assumed that meant I was good at dealing with difficult cases, and must have read a bunch of books about how to defuse crises. I did nothing to disabuse them of this.

Then a few days ago I stumbled across the Reddit thread Has Anyone Here Ever Been To An LW/SSC Meetup Or Otherwise Met A Rationalist IRL? User dgerard wrote about

meeting me in 2011, saying:

> His superpower is that he projects a Niceness Field, where people talking to him face to face want to be more polite and civil. The only person I've met with a similar Niceness Field is Jimmy Wales from Wikipedia…when people are around [Jimmy] talking to him they feel a sort of urge to be civil and polite in discourse 🙂 I've seen people visibly trying to be very precise and polite talking to him about stuff even when they're quite upset about whatever it is. Scott has this too. It's an interesting superpower to observe.

I should admit nobody else has mentioned anything like this, and that narcissism biases me toward believing anyone who says I have a superpower. Still, it would explain a lot. And not necessarily in a good way. I've always believed psychodynamic therapies are mostly ineffective, and cognitive-behavioral therapies very effective, because all my patients seem to defy the psychodynamic mode of having having weird but emotionally dramatic reactions to things in their past, but conform effortlessly to the cognitive-behavioral mode of being able to understand and rationally discuss their problems. And the more I examine this, the more I realize that my results are pretty atypical for psychiatrists. There's something I'm doing – totally by accident – to produce those results. This is worrying not just as a psychiatrist, but as someone who wants to know anything about other people at all.

II.

New topic: paranoia and Williams Syndrome.

Paranoia is a common symptom of various psychiatric disorders – most famously schizophrenia, but also paranoid personality disorder, delusional disorder, sometimes bipolar disorder. You can also get it from abusing certain drugs – marijuana, LSD, cocaine, and even prescription drugs like Adderall and Ritalin. The fun thing about paranoia is how gradual it is. Sure, if you abuse every single drug at once you'll think the CIA is after you with their mind-lasers. But if you just take a *little* more Adderall than you were supposed to, you'll be 1% paranoid. You'll have a very mild tendency to interpret ambiguous social signals just a little bit more negatively than usual. If a friend leaves without saying goodbye, and you would normally think "Oh, I guess she had a train to catch", instead you think "Hm, I wonder what she meant by that". There are a bunch of good stimulant abuse cases in the literature that present as "patient's boss said she was unusually standoffish and wanted her to get psychiatric evaluation", show up in the office as "well of course I'm standoffish, everyone in my office excludes me from everything and is rude in a thousand little ways throughout the day", and end up as "cut your Adderall dosage in half, please".

("Why is that psychiatrist telling me to cut my Adderall in half? Does he think I'm lying about having ADHD? Is he calling me a liar? These doctors have always treated me like garbage. I HAVE RIGHTS, YOU KNOW!")

Williams Syndrome is much rarer – only about 1/10,000 people, and most of them die before reaching adulthood. It's marked by a sort of anti-paranoia; Williams patients are incapable of distrusting anyone. NPR has a good article, A Life Without Fear, describing some of what they go through:

> Kids and adults with Williams love people, and they are literally pathologically trusting. They have no social fear. Researchers theorize that this is probably because of a problem in their limbic system, the part of the brain that regulates emotion. There appears to be a disregulation in one of the chemicals (oxytocin)

> that signals when to trust and when to distrust. This means that it is essentially biologically impossible for [them] to distrust.

The results are less than heartwarming:

> As Isabelle got older, the negative side of her trusting nature began to play a larger role. A typical example happened a couple of years ago, when Jessica and her family were spending the day at the beach. Isabelle had been begging Jessica to go to Dairy Queen, and Jessica had been putting her off. Then Isabelle overheard a lady just down the beach.
>
> “She was telling her kids, ‘OK, let’s go to the Dairy Queen,’ ” Jessica says. “And so Isabelle went over and got into the lady’s van, got in the back seat, buckled up and was waiting to be taken to Dairy Queen with that family.”
>
> Jessica had no idea what had happened to Isabelle and was frantically searching for her when the driver of the van approached her and explained that she had been starting her car when she looked up and saw Isabelle’s face in the rearview mirror.
>
> The woman, Jessica says, was incredibly angry.
>
> “She said, ‘I am a stranger, you know!’ ” Jessica says. Essentially, the woman blamed Jessica for not keeping closer watch on her daughter — for neglecting to teach her the importance of not getting into a car with someone she didn’t know. But the reality could not be more different. “It’s like, ‘My friend, you have no idea,’ ” Jessica says.
>
> In fact, because of Isabelle, Jessica has had to rethink even the most basic elements of her day-to-day life. She can not take Isabelle to the dog park. She tries not to take Isabelle to the store. And when the doorbell rings, Jessica will leap over a coffee table to intercept her.
>
> It’s not just Jessica and her family who must be vigilant. Every teacher at Isabelle’s public school has been warned. Isabelle is not allowed to tell them that she loves them. Isabelle is not supposed to tell other schoolchildren that she loves them. And there are other restrictions.
>
> “She’s not allowed to go to the bathroom alone at her school, because there have been numerous instances of girls with Williams syndrome being molested at school when they were alone in the hallway,” Jessica says. “And these are like middle class type schools. So it’s a very real problem. And, you know, I’d rather her be overly safe than be on CNN.”

Some of the research on these kids is fascinating – I’m not sure I believe the study finding that they’re incapable of racism, but the one finding a deficit detecting anger in faces seems pretty plausible.

Williams Syndrome usually involves mental retardation, but not always. Some of these people have normal IQ. It doesn’t really help. Threat-detection seems to be an automated process not totally susceptible to System II control. Maybe it’s like face-blindness. Intelligence can help a face-blind person come up with some systems to reduce the impact of their condition, but in the end it’s just not going to help that much.

Psychiatric disorders are often at the extremes of natural variation in human traits. For every intellectually disabled person, there are a dozen who are just kind of dumb. For every autistic person, there are a dozen who are just sort of nerdy. And so on. We naturally think of some people as more trusting than others, but maybe that isn't the best frame. "Trusting" implies that we all receive the same information, and just choose how much risk we're willing to tolerate. I don't know if that's true at all.

A recent theme here has been the ways that our sense-data is underdetermined. Each datum permits multiple possible explanations: this is true of visual and auditory perception, but also of the social world. A pretty girl laughs a little too long at a man's joke; is she trying to flirt with him, or just friendly? A boss calls her subordinate's work "okay" – did she mean to compliment him, or imply it was mediocre? A friend breaks off two appointments in a row, each time saying that something has come up – did something come up, or is he getting tired of the friendship? These are the sorts of questions everyone navigates all the time, usually with enough success that when autistic people screw them up, the rest of society nods sagely and says they need to learn to understand how to read context.

But "context" means "priors", and priors can differ from person to person. There's a lot of room for variation here before we get to the point where somebody will be so off-base that they end up excluded from society. Just as there's a spectrum from smart to dumb, or from introverted to extraverted, so there's a spectrum in people's tendencies to interpret ambiguous situations in a positive or negative way. There are people walking around who are just short of clinically paranoid, or just shy of Williams Syndrome levels of trust. And this isn't a value difference, it's a perceptual one. These people aren't bitter or risk-averse – or at least they don't *start off* that way. They just notice how everyone's hostile to them, all the time.

III.

Another change in topic: bubbles.

I've written before about how 46% of Americans are young-earth creationists, and how strongly that fails to square with my personal experience. I've met young-earth creationists once or twice. But of my hundred closest friends/co-workers/acquaintances, I think zero percent of them fall in that category. I'm not intentionally selecting friends on the basis of politics, religion, or anything else. It just seems to have happened. Something about my personality, location, social class, et cetera has completely isolated me from one particular half of the US population; I'm living in a non-creationist bubble in the midst of a half-creationist country.

What other bubbles do I live in? A quick look over my Facebook and some SSC survey results finds that my friends are about twenty times more likely to be transgender than the general population. There are about twice as many Asians but less than half as many African-Americans. Rates of depression, OCD, and autism are sky-high; rates of drug addiction and alcoholism are very low. Programmers are overrepresented at about ten times the Bay Area average.

I didn't intend any of these bubbles. For example, I've never done any programming myself, I'm not interested in it, and I try my best to avoid programmer-heavy places where I know all the conversations are going to be programming-related. Hasn't helped. And I'm about as cisgender as can be, I have several Problematic opinions, and I still can't keep track of which gender all of my various friends are on a month-to-month basis. Part of it is probably class-, race-, and location-based. And I have some

speculative theories about the rest – I think I have a pretty thing-oriented/systematizing thinking style, and so probably I get along better with other groups disproportionately made up of people whose thoughts work the same way – but I didn't understand any of this until a few years ago and there are still some parts that don't make sense. For now I just have to accept it as a given.

There are other bubbles I understand much better. Most of my friends are pretty chill and conflict-averse. This is because I used to have scarier conflict-prone friends, and as soon as I got into conflicts with them, I broke off the friendship. I'm not super-proud of this and it's probably one of those maladaptive coping styles you always hear about, and a lot of people have told me I'm really extreme on this axis and need to be better at tolerating aggressive people – but whenever I try, I find it unpleasant and stop. I know some other people who seem to actively seek out abrasive types so they can get in fun fights with them. I don't understand these people at all – but whatever their thought processes, we have different bubbles.

All of this goes double or triple for people I've dated. I don't think of myself as clearly having a "type", but people I date tend to turn out similar in dimensions I didn't expect when I first met them. I'm going to be ambiguous here because it's a small enough sample that I don't want to give away people's private information, but it's true.

I think about this a lot when I meet serial abuse victims.

These people are a heartbreaking psychiatric cliche. Abused by their parents, abused by their high school boyfriend, abused by their first husband, abused by their second husband, abused by the guy they cheated on their first husband with, abused by the friend they tried to go to for help dealing with all the abuse. The classic (though super offensive) explanation is that some people seek out abusers for some reason – maybe because they were abused as children and they've internalized that as the "correct" model of a relationship.

And maybe this is true for some people. I have a friend who admits it's true of her – her current strategy is to try to find someone in the sweet spot between "jerkish/narcissistic enough to be interesting" and "jerkish/narcissistic enough to actually abuse her", and she's said so in so many words to people trying to matchmake. I guess all I can do is wish her luck.

But for a lot of people, this sort of claim is just as offensively wrong as it sounds. I know people who have tried really hard to avoid abusers, who have gone to therapy and asked their therapist for independent verification that their new partner doesn't seem like the abusive type, who have pulled out all the stops – and who still end up with abusive new partners. These people are cursed through no fault of their own. All I can say is that whatever mysterious forces connect me to transgender pro-evolution programmers are connecting them to abusers. Something completely unintentional that they try their best to resist gives them a bubble of terrible people.

I want to emphasize as hard as I can that I'm not blaming them or saying there's anything they can do about their situation, and I have no doubt that despite my emphasis people are still going to accuse me of saying this, and I apologize if any of this sounds at all like anything in this direction. But *something* has to be happening here.

IV.

Sometimes I write about discrimination, and people send me emails about their own experiences. Many sound like this real one (quoted here with permission) from a woman who studied computer science at MIT and now works in the tech industry:

> In my life, I have never been catcalled, inappropriately hit on, body-shamed, unwantedly touched in a sexual way, discouraged from a male-dominated field, told I couldn't do something because it was a boy thing, or suffered from many other experiences that have traditionally served as examples as ways that women are less privileged. I have also never been shamed for not following gender norms (e.g. doing a bunch of math/science/CS stuff); instead I get encouraged and told that I'm a role model. I've never had problems going around wearing no make-up, a t-shirt, and cargo pants; but on the rare occasion that I do wear make-up / wear a dress, that's completely socially acceptable…Hopefully my thoughts/experiences are helpful for your future social justice based discussions.

Other times they sound like the opposite. I don't have anyone in this category who's given me permission to quote their email verbatim (consider ways this might not be a coincidence), but they're pretty much what you'd expect – a litany of constantly being put down, discriminated against, harassed, et cetera, across multiple jobs, at multiple companies, to the point where they complain it's "endemic" (I guess I can quote one word) and that we need to reject a narrative of "a few bad apples" because really it's a problem with *all* men to one degree or another.

These dueling categories of emails have always confused me. At the risk of being exactly the sort of creepy person the second set of writers complain about, I hunted down some of these people's Facebook profiles to see if one group was consistently more attractive than the other. They weren't. Nor is there any clear pattern in what industries or companies they work at, what position they're in, or anything else like that. There isn't even a consistent pattern in their politics. The woman I quote above mentions that she's a feminist who believes discrimination is a major problem – which has only made it extra confusing to her that she never experiences any of it personally.

These people don't just show up in my inbox. Some of them write articles on Slate, Medium, even The New Yorker, discussing not just how they've never experienced discrimination, but how much anger and backlash they've received when they try to explain this to everyone else. And all of them acknowledge that they know other people whose experiences seem to be the direct opposite.

I used to think this was pretty much just luck of the draw – some people will end up with nice people at great companies, other people will end up with bigots at terrible companies. I no longer think this explains everybody. Take that New Yorker article, by a black person who grew up in the South and says she was never discriminated against even once. I assume in her childhood she met thousands of different white Southerners; that's a pretty big lucky streak for none of them at all to be racists, especially when you consider all the people who report daily or near-daily harassment. Likewise, when you study computer science in college and then work in half a dozen tech companies over the space of decades and never encounter one sexist, that's quite the record. Surely something else must be going on here.

V.

And I think this has to come back to the sorts of things discussed in Parts I, II, and III.

People self-select into bubbles along all sorts of axes. Some of these bubbles are obvious and easy to explain, like rich people mostly meeting other rich people at the country club. Others are more mysterious, like how some non-programmer ends up with mostly programmer friends. Still others are horrible and completely outside comprehension, like someone who tries very hard to avoid abusers but ends up in multiple abusive relationships anyway. Even for two people living in the same country, city, and neighborhood, they can have a "society" made up of very different types of people.

People vary widely on the way they perceive social interaction. A paranoid schizophrenic will view every interaction as hostile; a Williams Syndrome kid will view every interaction as friendly. In between, there will be a whole range of healthy people without any psychiatric disorder who tend toward one side or the other. Only the most blatant data can be interpreted absent the priors that these dispositions provide; everything else will only get processed through preexisting assumptions about how people tend to act. Since things like racism rarely take the form of someone going up to you and saying "Hello, I am a racist and because of your skin color I plan to discriminate against you in the following ways…", they'll end up as ambiguous stimuli that everyone will interpret differently.

Finally, some people have personalities or styles of social interaction that unconsciously compel a certain response from their listeners. Call these "niceness fields" or "meanness fields" or whatever: some people are the sort who – if they became psychotherapists – would have patients who constantly suffered dramatic emotional meltdowns, and others' patients would calmly discuss their problems.

The old question goes: are people basically good or basically evil? Different philosophers give different answers. But so do different random people I know who aren't thinking philosophically at all. Some people describe a world of backstabbing Machiavellians, where everybody's a shallow social climber who will kick down anyone it takes to get to the top. Other people describe a world where everyone is basically on the same page, trying to be nice to everyone else but getting stuck in communication difficulties and honest disagreements over values.

I think both groups are right. Some people experience worlds of basically-good people who treat them nicely. Other people experience worlds of awful hypocritical backstabbers. This can be true even if they live in the same area as each other, work the same job as each other, et cetera.

And it's not just a basic good-evil axis. It can be about whether people are emotional/dramatic or calm/rational. It can be about whether people almost always discriminate or almost never do. It can be about whether they're honest or liars, shun outsiders or accept them, welcome criticism or reject it. Some people think elites are incompetent parasites; others that they're shockingly competent people who mean well and have interesting personalities. Some people think Silicon Valley is full of overpriced juicers, other people that it's full of structured-light engines. And the people who say all these things are usually accurately reporting their own experiences.

Some people are vaguely aware of this in the form of "privilege", which acknowledges different experiences at the cost of saying they have to line up exactly along special identity categories like race and gender. These certainly don't help, but it's not that simple – as proven by the article by that black Southerner who says she never once encountered discrimination. I've seen completely incomprehensible claims about

human nature by people of precisely the same race, sex, class, orientation, etc as myself, and I have no doubt they're trying to be truthful. The things that divide us are harder to see than we naively expect. Sometimes they're completely invisible.

To return to a common theme: *nothing makes sense except in light of inter-individual variation*. Variation in people's internal experience. Variation in people's basic beliefs and assumptions. Variation in level of abstract thought. And to all of this I would add a variation in our experience of other people. Some of us are convinced, with reason, that humankind is basically good. Others start the day the same way Marcus Aurelius did:

> When you wake up in the morning, tell yourself: the people I deal with today will be meddling, ungrateful, arrogant, dishonest, jealous and surly. They are like this because they cannot tell good from evil.

Notice this distinction, this way in which geographic neighbors can live in different worlds, and other people's thoughts and behaviors get a little more comprehensible.

Tech vs. Willpower

This is a linkpost for http://nautil.us/issue/52/the-hive/modern-media-is-a-dos-attack-on-your-free-will

Key pull quote, to me:

> *In your essay, you argue that the way these technologies indulge our impulsive selves breaks three kinds of attention necessary for democracy. What are they?*
>
> This is more a heuristic that I use. It's not a scientific argument. First, the "spotlight" of attention is how cognitive scientists tend to talk about perceptual attention. The things that are task-salient in my environment. How I select and interact with those, basically. Second, the "starlight." If the spotlight is about doing things, the starlight is who I want to be, not just what I want to do. It's like those goals that are valuable for their own sake, not because they're instrumental toward some other goal. Also, over time, how we keep moving toward those, and how we keep seeing the connections between the tasks we're doing right now, and those higher-level or longer-term goals. Third, the "daylight." In the philosopher Harry Frankfurt's terms, it's wanting what you want to want—the domain of metacognition. Basically, if the "spotlight" and the "starlight" are about pursuing some goal, some end, some value, the "daylight" is about the capacities that enable us to discern and define what those goals, those ends, are to begin with.

The full essay this references is not yet available, but there are extracts here

Infant Mortality and the Argument from Life History

Many people argue that suffering predominates in nature. A really simple form of the argument, supported by people like Brian Tomasik, is what one might call the argument from life history. In general, in most species, females produce many more offspring than can survive to adulthood; in some cases, a female may produce thousands or millions of offspring in a single reproductive season. Therefore, one can assume that most animals die before they are able to reproduce. In many cases, the offspring die before they can reasonably be considered conscious (for instance, an egg is eaten shortly after laying). However, even if half of animals die unconscious, the other half are a large source of disutility. Since death is generally quite painful, they may not have had enough positive experiences to outweigh the extraordinarily negative experience of death. It can therefore be assumed that there is more suffering than happiness in nature.

While this argument is intuitively compelling, I am not sure that it accurately reflects most people's opinions about how happiness works, so I have decided to write up three thought experiments that might help people think about it. These thought experiments are quite preliminary; I hope to spark a discussion so that people who are concerned about wild-animal suffering can debate.

1. The Human History Thought Experiment

Although the human population has been growing for thousands of years, for most of history the growth was fairly slow, suggesting the argument from life history applies to us as well. In part, that was because many humans died in childhood: for example, in 1800 four-tenths of humans died before they were five years old, a quarter of humans before their first birthday. (Note that 1800 is fairly late, and the statistics may have been even more stark in, say, 1 CE.)

I do not mean to deny that pre-modern human life was miserable in many ways: people were hungry, diseased, and poor. And I certainly don't mean to claim that high child mortality rates weren't a tragedy. However, my intuition is that-- whether or not human lives were worth living before modernity-- the high child mortality rate does not single-handedly prove that human lives were not worth living. Other information must be gathered to prove that. I suspect many other people's intuitions will agree.

To the extent the human history argument is misleading or anthropomorphizing, it strengthens my point: for instance, humans grieve their infants, while fish do not, so high infant mortality rates are worse for humans than for fish.

2. The Babykillers Thought Experiment

Humans have relatively few children and are growing (albeit slowly), perhaps suggesting that they are misleading as a thought experiment. Consider, therefore, a sapient species of aliens, the Babykillers. This species lays a thousand live young at a time. The young devour each other and only the single strongest offspring survives. All of the non-surviving thousand have miserable lives: they are tremendously hungry until they are eaten alive. The Babykillers have no way of modifying themselves to lay only a single offspring. To set aside issues of the Babykillers being replaced by

humans, assume that Babykillers are not aware of any other species. Would it be ethically required to have a Voluntary Babykiller Extinction Movement?

Personally, I put some weight on the argument that diversity is intrinsically valuable and therefore it is harmful to eliminate a sapient species. Otherwise, my moral intuitions are conflicted about whether Babykillers are net-negative and should be extinct.

3. The Long-Lived Babykillers Thought Experiment

The Babykillers are a species as specified above, except that the Babykillers are also extraordinarily long-lived: the average Babykiller who survives infancy lives for a thousand years. The average Babykiller who is eaten lives for only an hour. Babykillers are also a happy and fulfilled species. Therefore, while 999/1000 Babykillers experience a short life of great hunger followed by a painful death, only about one in ten thousand hours experienced by a Babykiller consists of great hunger and a painful death. The rest are quite happy.

The Long-Lived Babykiller thought experiment is supposed to get at an alternate method of assessing well-being. Instead of thinking about whether the average member of a species has a life worth living, we instead think about whether the average hour experienced by a member of a species is worth experiencing. For species with high juvenile mortality and/or long lives, these may be very different metrics.

Intuitively, I think the Long-Lived Babykillers should not go extinct. I think that also goes along with my intuitions about the human case. Since humans are fairly long-lived, a relatively small percentage of human hours are spent being a sick infant. I'm also tempted by the practical benefits. You could, in theory, figure out whether an animal species's existence is net positive simply by randomly sampling animals (although of course seasonal changes such as winter starvation or mating seasons would make this more complicated). However, many people have prioritarian intuitions. In that case, the experiences of animals who die young and painfully should be given more weight than the experiences of happy animals.

To be clear, I'm not necessarily satisfied by the "average hour" criterion. I do think we haven't put enough philosophical work into understanding what makes a species's existence net positive or net negative, and I hope my thought experiments will prompt some thought about the issue.

HOWTO: Screw Up The LessWrong Survey and Bring Great Shame To Your Family

This is a linkpost for http://lesswrong.com/r/discussion/lw/ph4/howto_screw_up_the_lesswrong_survey_and_bring/

Seek Fair Expectations of Others' Models

Epistemic Status: Especially about the future.

Response To (Eliezer Yudkowsky): There's No Fire Alarm for Artificial General Intelligence

It's long, but read the whole thing. Eliezer makes classic Eliezer points in classic Eliezer style. Even if you mostly know this already, there's new points and it's worth a refresher. I fully endorse his central point, and most of his supporting arguments.

What Eliezer has rarely been, is fair. That's part of what makes The Sequences work. I want to dive in where he says he's going to be blunt – as if he's ever not been – so you know it's gonna be good:

> Okay, let's be blunt here. I don't think most of the discourse about AGI being far away (*or* that it's near) is being generated by models of future progress in machine learning. I don't think we're looking at wrong models; I think we're looking at no models.
>
> I was once at a conference where there was a panel full of famous AI luminaries, and most of the luminaries were nodding and agreeing with each other that of course AGI was very far off, except for two famous AI luminaries who stayed quiet and let others take the microphone.
>
> I got up in Q&A and said, "Okay, you've all told us that progress won't be all that fast. But let's be more concrete and specific. I'd like to know what's the *least* impressive accomplishment that you are very confident *cannot* be done in the next two years."
>
> There was a silence.
>
> Eventually, two people on the panel ventured replies, spoken in a rather more tentative tone than they'd been using to pronounce that AGI was decades out. They named "A robot puts away the dishes from a dishwasher without breaking them", and Winograd schemas. Specifically, "I feel quite confident that the Winograd schemas—where we recently had a result that was in the 50, 60% range—in the next two years, we will not get 80, 90% on that regardless of the techniques people use."
>
> A few months after that panel, there was unexpectedly a big breakthrough on Winograd schemas. The breakthrough didn't crack 80%, so three cheers for wide credibility intervals with error margin, but I expect the predictor might be feeling slightly more nervous now with one year left to go. (I don't think it was the breakthrough I remember reading about, but Rob turned up this paper as an example of one that could have been submitted at most 44 days after the above conference and gets up to 70%.)
>
> But that's not the point. The point is the silence that fell after my question, and that eventually I only got two replies, spoken in tentative tones. When I asked for concrete feats that were impossible in the next two years, I think that that's when

> the luminaries on that panel switched to trying to build a mental model of future progress in machine learning, asking themselves what they could or couldn't predict, what they knew or didn't know. And to their credit, most of them did know their profession well enough to realize that forecasting future boundaries around a rapidly moving field is actually *really hard*, that nobody knows what will appear on arXiv next month, and that they needed to put wide credibility intervals with very generous upper bounds on how much progress might take place twenty-four months' worth of arXiv papers later.
>
> (Also, Demis Hassabis was present, so they all knew that if they named something insufficiently impossible, Demis would have DeepMind go and do it.)
>
> The question I asked was in a completely different genre from the panel discussion, requiring a mental context switch: the assembled luminaries actually had to try to consult their rough, scarce-formed intuitive models of progress in machine learning and figure out what future experiences, if any, their model of the field definitely prohibited within a two-year time horizon. Instead of, well, emitting socially desirable verbal behavior meant to kill that darned hype about AGI and get some predictable applause from the audience.
>
> I'll be blunt: I don't think the confident long-termism has been thought out at all. If your model has the extraordinary power to say what will be impossible in ten years after another one hundred and twenty months of arXiv papers, then you ought to be able to say much weaker things that are impossible in two years, and you should have those predictions queued up and ready to go rather than falling into nervous silence after being asked.
>
> In reality, the two-year problem is hard and the ten-year problem is laughably hard. The future is hard to predict in general, our predictive grasp on a rapidly changing and advancing field of science and engineering is very weak indeed, and it doesn't permit narrow credible intervals on what can't be done.

I agree that most discourse around AGI is not based around models of machine learning. I agree the AI luminaries seem to not have given good reasons for their belief in AGI being far away.

I also think Eliezer's take on their response is entirely unfair. Eliezer asks an excellent question, but the response is quite reasonable.

I

It is *entirely* unfair to expect a queued up answer.

Suppose I have a perfectly detailed mental model for future AI developments. If you ask, "What's the chance ML can put away the dishes within two years?" I'll need to do math, but: 3.74%.

Eliezer asks me his question.

Have I recently worked through that question? There are tons of questions. Questions about least impressive things in any reference class are rare. Let alone this particular class, confidence level and length of time.

So, no. Not queued up. The only reason to have this answer queued up is *if someone is going to ask.*

I did not anticipate that. I certainly did not *in the context of a listening Dennis Hassabis.* This is quite the isolated demand for rigor*.* I'll need to think.

II

Assume a mental model of AI development.

I am asked for the *least impressive* thing. To answer well, I must maximize.

What must be considered?

I need to decide what Eliezer meant by very confident, and what *other people* will think it means, and what they think Eliezer meant. Three different values. Very confident as actually used varies wildly. Sometimes it means 90% or less. Sometimes it means 99% or more. Eliezer later claims I should know what my model *definitely prohibits* but asked about *very confident*. There is danger of misinterpretation.

I need to decide what impressiveness means in context. Impressiveness in terms of currently perceived difficulty? In terms of the public or other researchers going 'oh, cool'? Impressive for a child? Some mix? Presumably Eliezer means perceived difficulty but there is danger of willful misinterpretation.

I need to query my model slash brainstorm for unimpressive things I am very confident cannot be done in two years. I adjust for the Hassabis effect that tasks I name will be accomplished faster.

I find the least impressive thing.

Finally I choose whether to answer.

This process isn't fast *even with a full model of future AI progress.*

III

I have my answer: "A robot puts away the dishes from a dishwasher without breaking them."

Should I say it?

My upside is limited.

It won't be the least impressive thing not done within two years. Plenty of less impressive things *might* be done within two years. Some will and some won't. My answer will seem lousy. The Hassabis effect compounds this, since some things that did not happen in two years *might have if I'd named them.*

Did Eliezer's essay accelerate work done on unloading a dishwasher? On the Winograd schemas?

If I say something that *doesn't* happen but comes *close*, such as getting 80% on the Winograd schemas if we get to 78%, I look wrong and lucky. If it *doesn't* come close, I look foolish.

Also, humans are terrible at calibration.

A true 98% confident answer looks *hopelessly* conservative to most people, and my off-the-cuff 98% confident answer likely isn't 98% reliable.

Whatever I name might happen. How embarrassing! People will laugh, distrust and panic. My reputation suffers.

The answer Eliezer gets might be important. If I don't want laughter, distrust or panic, it might be bad if *even one* answer given happens within two years.

In exchange, Eliezer sees a greater willingness to answer, and I transfer intuition. Does that seem worth it?

IV

Eliezer asked his question. What happened?

The room fell silent. Multiple luminaries *stopped to think.* That seems excellent. Positive reinforcement!

Two gave tentative answers. Those answers seemed honest, reasonable and interesting. The question was *hard.* They were on the spot. Tentativeness was the opposite of a missing mood. It properly expresses low confidence. Positive reinforcement!

Others chose not to answer. Under the circumstances, I sympathize.

These actions do not seem like strong evidence of a lack of models, or of bad faith. This seems like what you hope to see.

V

I endorse Eliezer's central points. There will be no fire alarm. We won't have a clear sign AGI is coming soon until AGI arrives. We need to act *now.* It's an emergency *now.* Public discussion is mostly not based on models of AI progress or concrete short term predictions.

Most discussions of the future are not built around concrete models of the future. It is unsurprising that AI discussions follow this pattern.

One can still challenge that one needs short-term predictions about AI progress to make long-term predictions. It is *not obvious* long-term prediction is *harder*, or that it *depends upon* short-term predictions. AGI might come purely from incremental machine learning progress. It might require major insights. It might not come from machine learning.

There are many ways to then conclude that AGI is far away where far away means decades out. Not that decades out is all that far away. *Eliezer conflating the two should freak you out.* AGI reliably forty years away would be quite the fire alarm.

You could think there isn't much machine learning progress, or that progress is nearing its limits. You could think that progress will slow dramatically, perhaps because problems will get exponentially harder.

You might think problems will get exponentially harder and resources spent will get exponentially larger too, so estimates of future progress move mostly insofar as they move the expected growth rate of future invested resources.

You could think incentive gradients from building more profitable or higher scoring AIs won't lead to AGIs, even if other machine learning paths might work. Dario Amodei says OpenAI is "following the gradient."

You could believe our civilization incapable of effort that does not follow incentive gradients.

You might think that our civilization will collapse or cease to do such research before it gets to AGI.

You could think building an AGI would require doing a thing, and our civilization is no longer capable of doing things.

You could think that there is a lot of machine learning progress to be made between here and AGI, such that even upper bounds on current progress leave decades to go.

You could think that even a lot of the right machine learning progress won't lead to AGI at all. Perhaps it is an entirely different type of thought. Perhaps it does not qualify as thought at all. We find more and more practical tasks that AIs can do with machine learning, but one can think both 'there are a lot of tasks machine learning will learn to do' and 'machine learning in anything like its current form cannot, even fully developed, do all tasks needed for AGI.'

And so on.

Most of those don't predict much about the next two years, other than a non-binding upper bound. With these models, when machine learning does a new thing, that teaches us more about that problem's difficulty than about how fast machine learning is advancing.

Under these models, Go and Heads Up No-Limit Hold 'Em Poker are easier problems than we expected. We should update in favor of well-defined adversarial problems with compact state expressions but large branch trees being easier to solve. That doesn't mean we shouldn't update our progress estimates at all, but perhaps we shouldn't update much.

This goes with everything AI learns to do ceasing to be AI.

Thus, one can reasonably have a model where impressiveness of short-term advances does not much move our AGI timelines.

I saw an excellent double crux on AI timelines, good enough to update me dramatically on the value of double crux and greatly enrich my model of AI timelines. Two smart, highly invested people had given the problem a lot of thought, and were doing their best to build models and assign probabilities and seek truth. Many questions came up. Short-term concrete predictions did not come up. At all.

VI

That does *not* mean any of that is what is happening.

I think mostly what Eliezer thinks is happening, is happening. People's incentive gradients on short term questions say not to answer. People's incentive gradients on long term questions say to have AGI be decades out. That's mostly what they answer. Models might exist, but why let them change your answer? If you answer AGI is near

and it doesn't happen you look foolish. If you answer AGI is near and it happens, *who cares what you said?*

When asked a question, good thinkers generate as much model as they need. Less good thinkers, or the otherwise motivated, instead model of what it is in their interest to say.

Most people who say productive AI safety work cannot currently be done have not spent two hours thinking about what could currently be done. Again, that's true of all problems. Most people *never* spend two hours thinking about what could be done about *anything*. Ever. See Eliezer entire essential sequence (sequence Y).

That is how someone got so frustrated with getting people to actually think *about AI safety* that he decided it would be easier to get them to actually think *in general.*

To do that, it's important to be *totally unfair* to not thinking. Following incentive gradients and social queues and going around with inconsistent models and not trying things for even five minutes before declaring them impossible won't cut it and *that is totally not OK.*

He emphasizes nature not grading on a curve, and *fails everyone. Hard.* The Way isn't just A Thing, it's a necessary thing.

Then we realize that no, it's *way worse* than that. People are not only not following The Way. No one does the thing they are supposedly doing. The world is mad on a different level than inaccurate models without proper Bayesian updating and not stopping to think or try for five minutes once in their life let alone two hours. There are no models anywhere.

Fairness can't always be a thing. Trying to make it a thing where it isn't a thing tends to go quite badly.

Sometimes, though, you still need fairness. Without it groups can't get along. Without it you can't cooperate. Without it we treat thinking about a new and interesting question as evidence of a lack of thinking.

Holding everyone to heroic responsibility wins you few friends, influences few people and drives you insane.

VII

Where does that leave us? Besides the original takeaway that There Is No Fire Alarm For Artificial General Intelligence and we need to work on the problem now? And your periodic reminder that people are crazy and the world is mad?

Microfoundations are great, but some useful models don't have them. It would be great if everyone had probabilistic time distributions for every possible event, but this is *totally not reasonable,* and *totally not required to have a valid opinion.* Some approaches answer some questions but not others.

We must hold onto our high standards for ourselves and those who opt into them. For others, we must think about circumstance and incentive, and stop at 'tough, but fair.'

Predictions are valuable. They are hard to do well and socially expensive to do honestly. A culture of stating your probabilities upon request is good. Betting on your

beliefs is better. Part of that is understanding not everyone has thought through everything. And understanding adverse selection and bad social odds. And realizing sometimes best guesses would get taken too seriously, or commit people to things. Sometimes people need to speak tentatively. Or say “I don’t know.” Or say nothing.

Allies won’t always ponder what you’re pondering. They aren’t perfectly rigorous thinkers. They don’t think hard for two hours about your problem. They don’t often make extraordinary efforts.

Most of what they want will involve social reality and incentive gradients and muddled thinking. They’re doing it for the wrong reasons. They will often be unreliable and untrustworthy. They’re defecting *constantly.*

You go to war with the army you have.

We can’t afford to hold everyone to impossible standards. Even holding *ourselves* to impossible standards requires psychologically safe ways to do that.

When someone genuinely thinks, and offers real answers, cheer that. Especially answers against interest. They do the best they can. From another perspective *they could obviously do so much more,* but one thing at a time.

Giving them the right social incentive gradient, even in a small way, matters a lot.

Someone is doing their best to break through the incentive gradients of social reality.

We can work with that.

Time to Exit the Sandbox

This is a linkpost for http://squirrelinhell.blogspot.com/2017/10/time-to-exit-sandbox.html

Geeks, Mops, Sociopaths

This is a linkpost for https://meaningness.com/geeks-mops-sociopaths

What would convince you you'd won the lottery?

The latest (06 Oct 2017) Euromillion lottery numbers were 01 - 09 - 15 - 19 - 25 , with the "Lucky Stars " being 01 - 07.

Ha! Bet I convinced no-one about those numbers. The odds against 01 - 09 - 15 - 19 - 25 / 01 - 07 being the true lottery numbers are about 140 million to one, so I must have been a fool to think you'd believe me. The odds that I decided to lie is far more that one in a 140 million, so it's very unlikely those are the true numbers.

Wait a moment. Something is wrong here. That argument could be used against any set of numbers; and yet one of them was the real winning set. The issue is that probabilities are not being compared properly.

Let S=(01,09,15,19,25,01,07), let W(S) be the fact these numbers won the lottery, let L be the fact that I lied about the winning number, and L(S) be the fact that I lied and claimed S as the winning numbers.

So though P(L) is indeed much higher than P(W(S)), generally P(W(S)) will be higher than P(L(S)). That's because while W(S) gets penalised by all the other values S could have been, so does L(S).

Note the key word "generally". The sum of P(L(S')), across all S', is P(L); this means that for most S', P(L(S')) must be low, less than 140 million times smaller than P(L). But it's possible that for some values of S', it might be much higher. If I'd claimed that the winning numbers were 01 - 02 -03 - 04 - 05 / 06 - 07, then you might have cause to doubt me (after taking into account selection bias in me reporting it, the number of lotteries going on around the world, and so on). The connections with Komogorov complexity should be obvious at this point: the more complex the sequence, the less likely it is that a human being will select it disproportionately often (and you can't select *all* sequences disproportionately often).

What if I'd claimed that I'd won that lottery? That seems like a 01 - 02 -03 - 04 - 05 / 06 - 07 - style claim. I'm saying that the 01 - 09 - 15 - 19 - 25 / 01 - 07 did win (fair enough) but that these numbers are special to me because I selected them. Here you should be more sceptical. But what if I linked to a few articles that showed me with a gigantic novelty check? It would still seem more likely that I'd hacked and faked those than that I'd actually won... or would it?

Let's make it more personal: what if you yourself ended up winning the lottery? Or at least, if you'd got/been given a ticket, and a friend told you that you'd won. And then another. And then people started calling you up, and you looked up the information on website, and there were news crews around... Still, odds of 140 million to one. It's still more likely that people are doing elaborate practical jokes on you, or that you're in a simulation.

First of all, are you so sure of those odds on practical jokes? There are maybe a few hundred lotteries in the world, drawing regularly (not to mention the other games of chance). If the practical jokes are more likely, that means that every week, there must be thousands of very elaborate fake lottery winning jokes, including multiple people lying (possible), and faked news crews and websites (much less likely). Do you really

think that thousands of those happened every week (and that this almost never makes the news)?

If not, then you must be prepared to face the fact that you might actually be a winner. But see what this implies: that a few conversations with friends, a minor browse of the internet, and some news crew-looking people are enough to subjectively overturn odds of *140 million to one*.

In a sense, this isn't surprising: 140 million is roughly 2^27, so it should only take 27 bits of information to convince you of that. But those are 27 bits from a reliable source. Given the possibility for deception and manipulation, it still seems that you can believe incredibly unlikely specific events, given very small amounts of information.

What seems to be happening is that we have a background picture of the universe, built up by life experience, and we think we have a pretty reliable impression about things like how likely our friends are to be honest/pranksters/lazy pranksters/super-inspired super-hacking super-organised vicious pranksters. Our background picture of the universe has way more than 27 bits, so we're not comparing "I won the lottery" with "someone is pranking me", but with "someone is pranking me and major things I thought I knew about reality are wrong".

What about the simulation possibility? Now, I don't believe that the "probability of being in a simulation" is a meaningful concept. But let's pretend it is. We already know that we can't compare "probability of winning the lottery" with "probability of being in a simulation", but instead with "probability of being in a simulation and winning the lottery".

Now, that last doesn't seem too unlikely, given a simulation. But that's still not the valid comparison. It's "probability of winning the lottery + background experience of the universe" versus "probability of being in a simulation and winning the lottery + background experience of the universe". It's that background experience that weighs heavily here. Sure, a simulation might be more likely to simulate you wining the lottery, but would it be likely to have first constructed a more conventional life for you, and then have you winning the lottery?

Knowing that you are in a simulation only matters if that information is useful to your plans. So we're not even talking about "being in a simulation and winning the lottery + background experience of the universe", but all that plus "and the future will be radically different from my current highly materialistic view of reality". The more it looks like the laws of physics are being followed, the more you should expect them to continue to look as if they are followed. And winning a lottery, though unlikely, is not a radical departure from physics. So the simulation hypothesis doesn't get all that much of a boost from that fact.

A conclusion?

Tradition maintains that posts must be concluded by conclusions. This post doesn't really have any strong ones, but here are some thoughts:

- Even if the dream/prank/simulation/everything-you-know-is-wrong hypothesis is *overall* more likely than event E, what matters is the probability of dream

conditional on you seeing the evidence you've seen for E, and this can be much much smaller.

- It doesn't take much evidence for us to (correctly) believe *extremely* unlikely things.
- One of the reasons for that is that we have background knowledge of the universe that provides a lot of bits of knowledge in our favour.
- We're not very good at estimating these kinds of probabilities.
- If you have decent-seeming evidence that you won the lottery, you probably did: due to background knowledge, this evidence can be stronger than it seems.

Distinctions in Types of Thought

Epistemic status: speculative

For a while, I've had the intuition that current machine learning techniques, though powerful and useful, are simply not touching some of the functions of the human mind. But before I can really get at how to justify that intuition, I would have to start clarifying what different kinds of thinking there are. I'm going to be reinventing the wheel a bit here, not having read that much cognitive science, but I wanted to write down some of the distinctions that seem important, and trying to see whether they overlap. A lot of this is inspired by Dreyfus' Being-in-the-World. I'm also trying to think about the questions raised in the post "What are Intellect and Instinct?"

Effortful vs. Effortless

In English, we have different words for perceiving passively versus actively paying attention. To see vs. to look, to hear vs. to listen, to touch vs. to feel. To go looking for a sensation means exerting a sort of mental pressure; in other words, effort. William James, in his Principles of Psychology, said "Attention and effort, as we shall see later, are but two names for the same psychic fact." He says, in his famous introduction to attention, that

> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Localization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German.

Furthermore, James divides attention into two types:

> Passive, reflex, non-voluntary, effortless ; or Active and voluntary.

In other words, the mind is always selecting certain experiences or thoughts as more salient than others; but sometimes this is done in an automatic way, and sometimes it's effortful/voluntary/active. A fluent speaker of a language will automatically notice a grammatical error; a beginner will have to try effortfully to catch the error.

In the famous gorilla experiment where subjects instructed to count passes in a basketball game failed to notice a gorilla on the basketball field, "counting the passes" is paying effortful attention, while "noticing the gorilla" would be effortless or passive noticing.

Activity in "flow" (playing a musical piece by muscle memory, or spatially navigating one's own house) is effortless; activities one learns for the first time are effortful.

Oliver Sacks' case studies are full of stories that illustrate the importance of flow. People with motor disorders like Parkinson's can often dance or walk in rhythm to music, even when ordinary walking is difficult; people with memory problems sometimes can still recite verse; people who cannot speak can sometimes still sing. "Fluent" activities can remain undamaged when similar but more "deliberative" activities are lost.

The author of intellectualizing.net thinks about this in the context of being an autistic parent of an autistic son:

> Long ago somewhere I can't remember, I read a discussion of knowing what vs. knowing how. The author's thought experiment was about walking. Imagine walking with conscious planning, thinking consciously about each muscle and movement involved. Attempting to do this makes us terrible at walking.
>
> When I find myself struggling with social or motor skills, this is the feeling. My impression of my son is the same. Rather than trying something, playing, experimenting he wants the system first. First organize and analyze it, then carefully and cautiously we might try it.
>
> A simple example. There's a curriculum for writing called Handwriting Without Tears. Despite teaching himself to read when barely 2, my son refused to even try to write. Then someone showed him this curriculum in which letters are broken down into three named categories according to how you write them; and then each letter has numbered strokes to be done in sequence. Suddenly my son was interested in writing. He approached it by first memorizing the whole Handwriting Without Tears system, and only then was he willing to try to write. I believe this is not how most 3-year-olds work, but this is how he works.
>
> …
>
> One simple study ("Children with autism do not overimitate") had to do with children copying "unnecessary" or "silly" actions. Given a demonstration by an adult, autistic kids would edit out pointless steps in the demonstrated procedure. Think about what's required to do this: the procedure has to be reconstructed from first principles to edit the silly out. The autistic kids didn't take someone's word for it, they wanted to start over.

The author and his son learn skills by effortful conscious planning that most people learn by "picking up" or "osmosis" or "flow."

Most of the activity described by Heidegger's *Being and Time*, and Dreyfus' commentary *Being-In-The-World*, is effortless flow-state "skilled coping." Handling a familiar piece of equipment, like typing on a keyboard, is a prototypical example. You're not thinking about how to do it except when you're learning how for the first time, or if it breaks or becomes "disfluent" in some way. If I'm interpreting him correctly, I think Dreyfus would say that neurotypical adults spend most of their time, minute-by-minute, in an effortless flow state, punctuated by occasions when they have to plan, try hard, or figure something out.

William James would agree that voluntary attention occupies a minority of our time:

> There is no such thing as voluntary attention sustained for more than a few seconds at a time. What is called sustained voluntary attention is a repetition of successive efforts which bring back the topic to the mind.

(This echoes the standard advice in mindfulness meditation that you're not aiming for getting the longest possible period of uninterrupted focus, you're training the mental motion of returning focus from mind-wandering.)

Effortful attention can also be viewed as the cognitive capacities which stimulants improve. Reaction times shorten, and people distinguish and remember the stimuli in

front of them better.

It's important to note that not all focused attention is effortful attention. If you are playing a familiar piece on the piano, you're in a flow state, but you're still being "focused" in a sense; you're noticing the music more than you're noticing conversation in another room, you're playing this piece rather than any other, you're sitting uninterrupted at the piano rather than multitasking. Effortless flow can be extremely selective and hyper-focused (like playing the piano), just as much as it can be diffuse, responsive, and easily interruptible (like navigating a crowded room). It's not the size of your window of salience that distinguishes flow from effortful attention, it's the pressure that you apply to that window.

Psychologists often call effortful attention cognitive disfluency, and find that experiences of disfluency (such as a difficult-to-read font) improve syllogistic reasoning and reduce reliance on heuristics, while making people more likely to make abstract generalizations. Disfluency improves results on measures of "careful thinking" like the Cognitive Reflection Test as well as on real-world high-school standardized tests, and also makes people less likely to confess embarrassing information on the internet. In other words, disfluency makes people "think before they act." Disfluency raises heart rate and blood pressure, just like exercise, and people report it as being difficult and reliably disprefer it to cognitive ease. The psychology research seems consistent with there being such a thing as "thinking hard." Effortful attention occupies a minority of our time, but it's prominent in the most specifically "intellectual" tasks, from solving formal problems on paper to making prudent personal decisions.

What does it mean, on a neurological or a computational level, to expend mental effort? What, precisely, are we doing when we "try hard"? I think it might be an open question.

Do the neural networks of today simulate an agent in a state of "effortless flow" or "effortful attention", or both or neither? My guess would be that deep neural nets and reinforcement learners are generally doing effortless flow, because they excel at the tasks that we generally do in a flow state (pattern recognition and motor learning.)

Explicit vs. Implicit

Dreyfus, as an opponent of the Representational Theory of Mind, believes that (most of) cognition is not only not based on a formal system, but not in principle formalizable. He thinks you couldn't possibly write down a theory or a set of rules that explain what you're doing when you drive a car, even if you had arbitrary amounts of information about the brain and human behavior and arbitrary amounts of time to analyze them.

This distinction seems to include the distinctions of "declarative vs. procedural knowledge", "know-what vs. know-how", savoir vs. connaître. We can often do, or recognize, things that we cannot explain.

I think this issue is related to the issue of interpretability in machine learning; the algorithm executes a behavior, but sometimes it seems difficult or impossible to explain what it's doing in terms of a model that's simpler than the whole algorithm itself.

The seminal 2001 article by Leo Breiman, "Statistical Modeling: The Two Cultures" and Peter Norvig's essay "On Chomsky and the Two Cultures of Statistical Learning" are

about this issue. The inverse square law of gravitation and an n-gram Markov model for predicting the next word in a sentence are both statistical models, in some sense; they allow you to predict the dependent variable given the independent variables. But the inverse square law is interpretable (it makes sense to humans) and explanatory (the variables in the model match up to distinct phenomena in reality, like masses and distances, and so the model is a relationship between things in the world.)

Modern machine learning models, like the n-gram predictor, have vast numbers of variables that don't make sense to humans and don't obviously correspond to things in the world. They perform well without being explanations. Statisticians tend to prefer parametric models (which are interpretable and sometimes explanatory) while machine-learning experts use a lot of non-parametric models, which are complex and opaque but often have better empirical performance. Critics of machine learning argue that a black-box model doesn't bring understanding, and so is the province of engineering rather than science. Defenders, like Norvig, flip open a random issue of Science and note that most of the articles are not discovering theories but noting observations and correlations. Machine learning is just another form of pattern recognition or "modeling the world", which constitutes the bulk of scientific work today.

These are heuristic descriptions; these essays don't make explicit how to test whether a model is interpretable or not. I think it probably has something to do with model size; is the model reducible to one with fewer parameters, or not? If you think about it that way, it's obvious that "irreducibly complex" models, of arbitrary size, can exist in principle — you can just build simulated data sets that fit them and can't be fit by anything simpler.

How much of human thought and behavior is "irreducible" in this way, resembling the huge black-box models of contemporary machine learning? Plausibly a lot. I'm convinced by the evidence that visual perception runs on something like convolutional neural nets, and I don't expect there to be "simpler" underlying laws. People accumulate a lot of data and feedback through life, much more than scientists ever do for an experiment, so they can "afford" to do as any good AI startup does, and eschew structured models for open-ended, non-insightful ones, compensating with an abundance of data.

Subject-Object vs. Relational

This is a concept in Dreyfus that I found fairly hard to pin down, but the distinction seems to be operating upon the world vs. relating to the world. When you are dealing with raw material — say you are a potter with a piece of clay — you think of yourself as active and the clay as passive. You have a goal (say, making a pot) and the clay has certain properties; how you act to achieve your goal depends on the clay's properties.

By contrast, if you're interacting with a person or an animal, or even just an object with a UI, like a stand mixer, you're relating to your environment. The stand mixer "lets you do" a small number of things — you can change attachments or speeds, raise the bowl up and down, remove the bowl, fill it with food or empty it. You orient to these affordances. You do not, in the ordinary process of using a stand mixer, think about whether you could use it as a step-stool or a weapon or a painting tool. (Though you might if you are a child, or an engineer or an artist.) Ordinarily you relate in an almost social, almost animist, way to the stand mixer. You use it as it "wants to be

used", or rather as its designer wants you to use it; you are "playing along" in some sense, being receptive to the external intentions you intuit.

And, of course, when we are relating to other people, we do much stranger and harder-to-describe things; we become different around them, we are no longer solitary agents pursuing purely internally-driven goals. There is such a thing as becoming "part of a group." There is the whole messy business of culture.

For the most part, I don't think machine-learning models today are able to do either subject-object or relational thinking; the problems they're solving are so simple that neither paradigm seems to apply. "Learn how to work a stand mixer" or "Figure out how to make a pot out of clay" both seem beyond the reach of any artificial intelligence we have today.

Aware vs. Unaware

This is the difference between sight and blindsight. It's been shown that we can act on the basis of information that we don't know we have. Some blind people are much better than chance at guessing where a visual stimulus is, even though they claim sincerely to be unable to see it. Being primed by a cue makes blindsight more accurate — in other words, you can have attention without awareness.

Anosognosia is another window into awareness; it is the phenomenon when disabled people are not aware of their deficits (which may be motor, sensory, speech-related, or memory-related.) In unilateral neglect, for instance, a stroke victim might be unaware that she has a left side of her body; she won't eat the left half of her plate, make up the left side of her face, etc. Sensations may still be possible on the left side, but she won't be aware of them. Squirting cold water in the left ear can temporarily fix this, for unknown reasons.

Awareness doesn't need to be explicit or declarative; we aren't formalizing words or systems constantly when we go through ordinary waking life. It also doesn't need to be effortful attention; we're still aware of the sights and sounds that enter our attention spontaneously.

Efference copy signals seem to provide a clue to what's going on in awareness. When we act (such as to move a limb), we produce an "efference copy" of what we expect our sensory experience to be, while simultaneously we receive the actual sensory feedback. "This process ultimately allows sensory reafferents from motor outputs to be recognized as self-generated and therefore not requiring further sensory or cognitive processing of the feedback they produce." This is what allows you to keep a 'still' picture of the world even though your eyes are constantly moving, and to tune out the sensations from your own movements and the sound of your own voice.

Schizophrenics may be experiencing a dysfunction of this self-monitoring system; they have "delusions of passivity or thought insertion" (believing that their movements or thoughts are controlled from outside) or "delusions of grandeur or reference" (believing that they control things with their minds that they couldn't possibly control, or that things in the outside world are "about" themselves when they aren't.) They have a problem distinguishing self-caused from externally-caused stimuli.

We're probably keeping track, somewhere in our minds, of things labeled as "me" and "not me" (my limbs are part of me, the table next to me is not), sensations that are self-caused and externally-caused, and maybe also experiences that we label as

“ours” vs. not (we remember them, they feel like they happened to us, we can attest to them, we believe they were real rather than fantasies.)

It might be as simple as just making a parallel copy of information labeled “self,” as the efference-copy theory has it. And (probably in a variety of complicated and as-yet-unknown ways), our brains treat things differently when they are tagged as “self” vs. “other.”

Maybe when experiences are tagged as “self” or labeled as memories, we are aware that they are happening to us. Maybe we have a “Cartesian theater” somewhere in our brain, through which all experiences we’re aware of pass, while the unconscious experiences can still affect our behavior directly. This is all speculation, though.

I’m pretty sure that current robots or ML systems don’t have any special distinction between experiences inside and outside of awareness, which means that for all practical purposes they’re always operating on blindsight.

Relationships and Corollaries

I think that, in order of the proportion of ordinary neurotypical adult life they take up, awareness > effortful attention > explicit systematic thought. When you look out the window of a train, you are aware of what you see, but not using effortful attention or thinking systematically. When you are mountain-climbing, you are using effortful attention, but not thinking systematically very much. When you are writing an essay or a proof, you are using effortful attention, and using systematic thought more, though perhaps not exclusively.

I think awareness, in humans, is necessary for effortful attention, and effortful attention is usually involved in systematic thought. (For example, notice how concentration and cognitive disfluency improve the ability to generalize or follow reasoning principles.) . I don’t know whether those necessary conditions hold in principle, but they seem to hold in practice.

Which means that, since present-day machine-learners aren’t aware, there’s reason to doubt that they’re going to be much good at what we’d call reasoning.

I don’t think classic planning algorithms “can reason” either; they’re hard-coding in the procedures they follow, rather than generating those procedures from simpler percepts the way we do. It seems like the same sort of misunderstanding as it would be to claim a camera can see.

(As I’ve said before, I don’t believe anything like “machines will never be able to think the way we do”, only that they’re not doing so now.)

The Weirdness of Thinking on Purpose

It’s popular these days to “debunk” the importance of the “intellect” side of “intellect vs. instinct” thinking. To point out that we aren’t always rational (true), are rarely thinking effortfully or explicitly (also true), can’t usually reduce our cognitive processes to formal systems (also true), and can be deeply affected by subconscious or subliminal processes (probably true).

Frequently, this debunking comes with a side order of sneer, whether at the defunct “Enlightenment” or “authoritarian high-modernist” notion that everything in the mind can be systematized, or at the process of abstract/deliberate thought itself and the

people who like it. Jonathan Haidt's lecture on "The Rationalist Delusion" is a good example of this kind of sneer.

The problem with the popular "debunking reason" frame is that it distracts us from noticing that the actual process of reasoning, as practiced by humans, is a phenomenon we don't understand very well yet. Sure, Descartes may have thought he had it all figured out, and he was wrong; but thinking still exists even after you have rejected naive rationalism, and it's a mistake to assume it's the "easy part" to understand. Deliberative thinking, I would guess, is the hard part; that's why the cognitive processes we understand best and can simulate best are the more "primitive" ones like sensory perception or motor learning.

I think it's probably better to think of those cognitive processes that distinguish humans from animals as weird and mysterious and special, as "higher-level" abilities, rather than irrelevant and vestigial "degenerate cases", which is how Heidegger seems to see them. Even if the "higher" cognitive functions occupy relatively little time in a typical day, they have outsize importance in making human life unique.

Two weirdly similar quotes:

> "Three quick breaths triggered the responses: he fell into the floating awareness… focusing the consciousness… aortal dilation… avoiding the unfocused mechanism of consciousness… **to be conscious by choice**… blood enriched and swift-flooding the overload regions… **one does not obtain food-safety freedom by instinct alone**… animal consciousness does not extend beyond the given moment nor into the idea that its victims may become extinct… the animal destroys and does not produce… animal pleasures remain close to sensation levels and avoid the perceptual… the human requires a background grid through which to see his universe… **focused consciousness by choice, this forms your grid**… bodily integrity follows nerve-blood flow according to the deepest awareness of cell needs… all things/cells/beings are impermanent… strive for flow-permanence within…"
>
> –Frank Herbert, Dune, 1965

> "An animal's consciousness functions automatically: an animal perceives what it is able to perceive and survives accordingly, no further than the perceptual level permits and no better. Man cannot survive on the perceptual level of his consciousness; his senses do not provide him with an automatic guidance, they do not give him the knowledge he needs, only the material of knowledge, which his mind has to integrate. Man is the only living species who has to perceive reality, which means: **to be conscious — by choice**. But he shares with other species the penalty for unconsciousness: destruction. For an animal, the question of survival is primarily physical; for man, primarily epistemological.
>
> "Man's unique reward, however, is that while animals survive by adjusting themselves to their background, man survives by adjusting his background to himself. If a drought strikes them, animals perish — man builds irrigation canals; if a flood strikes them, animals perish — man builds dams; if a carnivorous pack attacks them animals perish — man writes the Constitution of the United States. But **one does not obtain food, safety, or freedom — by instinct."**
>
> –Ayn Rand, For the New Intellectual, 1963

(bold emphasis added, ellipses original).

“Conscious by choice” seems to be pointing at the phenomenon of effortful attention, while “the unfocused mechanism of consciousness” is more like awareness. There seems to be some intuition here that effortful attention is related to the productive abilities of humanity, our ability to live in greater security and with greater thought for the future than animals do. We don’t usually “think on purpose”, but when we do, it matters a lot.

We should be thinking of “being conscious by choice” more as a sort of weird Bene Gesserit witchcraft than as either the default state or as an irrelevant aberration. It is neither the whole of cognition, nor is it unimportant — it is a special power, and we don’t know how it works.

Leaders of Men

Related to (Eliezer Yudkowsky): Inadequacy and Modesty

Epistemic Status: Confident. No sports knowledge required.

In 2005, Willie Randolph became manager of the New York Mets.

In his first five games as manager, all of which he lost, Willie made more decisions wrong than I thought possible. If he needed to change pitchers, he waited. Other times he changed pitchers for no reason. Starting lineups made zero sense. Position players bunted. And so on. He cost us *at least* one of those games. My friend Seth and called for Willie's head.

He would go on to an excellent 97 win season in 2006, come in second in manager-of-the-year voting, get a contract extension, and only get fired after wearing out our starting pitchers so much that we experienced one of the most epic late season collapses in baseball history in 2007, followed by a horrible 2008.

Willie's in-game decisions did not improve. If anything, they got worse.

Despite this, we came to understand why Willie got and kept his job.

Willie Randolph was a leader of men.

Players liked Willie. They wanted to play for him, work hard for him, be the best they could be. They put the team first. He created a positive clubhouse atmosphere. He inspired good performances, spotted ways players could improve.

That is what counts.

Do bad in-game decisions cost games? Absolutely. But not *that many* games. Maybe they lose you 4 a year out of 162.

If the lineup makes your players unhappy, that costs a lot more. If your pitchers lose motivation or have their rhythms disrupted, that matters more than getting high leverage for your best reliever. Maybe bunting inspires team unity. The reason we hate bunting so much isn't because it's a *huge* mistake. It's an *obvious* mistake.
A *pure* mistake. An arithmetic error.

Plenty of people could get those technical decisions right. I could do it.

What most of us *can't* do is lead men. Leading men *is what counts.* That's the real job, but it *comes with* these other tasks.

Sometimes these other tasks land in good hands, at other times they land in terrible hands. Those who do the little things right do succeed more, but you can still win championships without them. If you can lead men.

Other sports follow the pattern. Why do football teams employ Andy Reid, who could not manage a two-minute drill if his life depended on it? Why do highly successful basketball players refuse to cooperate with their teammates or practice key skills?

Because those are small mistakes. The things those people do right matter more.

Could Willie Randolph hire someone to micromanage the game? Could Andy Reid hire someone to manage his two minute drills?

No. The people who are capable of that, are not leaders of men, and how they make those decisions is part of how they lead men.

Even if you *could* do that, *fixing such penny-ante problems is too disruptive.* You want their eyes on the prize.

This generalizes.

If a position calls for a leader of men, you often find a leader of men. If it requires super high levels of another skill, whether it's coding, raising money, lifting weights, intricate chemistry or proving theorems, you'll find that. However, if you need *rare* levels of such skills compared to what you can offer, you won't select *for anything else.* You can't demand *ordinary competence* in insufficiently important areas. There aren't enough qualified applicants. Plus it wouldn't be worth the distraction.

This helps explain why people in unique positions are often *uniquely terrible.* They're not replaceable. Some incompetence and shenanigans are acceptable, so long as they deliver the goods.

The same goes for other groups, organizations, religions, software and most anything else.

If a system has *unique big advantages,* they're not effectively competing on less big things. They *might* be optimizing small things, but they don't have to, so you can't assume such things are optimized at all. Even when a system does not have unique advantages, anything *insufficiently central* is likely not optimized because it's not worthy of attention.

It is *much, much easier* to pick out a way in which a system is sub-optimal, than it is to implement or run that system at anything like its current level of optimization.

Thus I generally believe the following two things:

It is relatively easy to find ways in which almost anything could be improved on the margin, were one able to implement isolated changes. Well thought-out such ideas are often correct.

and also

The person making such a correct suggestion would likely be hopelessly lost trying to implement this change let alone running the relevant systems.

Yudkowsky on AGI ethics

A Cornell computer scientist recently wrote on social media:

> [...] I think the general sense in AI is that we don't know what will play out, but some of these possibilities are bad, and we need to start thinking about it. We are plagued by highly visible people ranging from Musk to Ng painting pictures ranging from imminent risk to highly premature needless fear, but that doesn't depict the center of gravity, which has noticeably shifted to thinking about the potential bad outcomes and what we might do about it. (Turning close to home to provide an example of how mainstream this is becoming, at Cornell two AI professors, Joe Halpern and Bart Selman, ran a seminar and course last semester on societal and ethical challenges for AI, and only just a few weeks ago we had a labor economist speak in our CS colloquium series about policy ideas targeting possible future directions for CS and AI, to an extremely large and enthusiastic audience.)

To which Eliezer Yudkowsky replied:

> My forecast of the net effects of "ethical" discussion is negative; I expect it to be a cheap, easy, attention-grabbing distraction from technical issues and technical thoughts that actually determine okay outcomes. [...]
>
> The ethics of bridge-building is to not have your bridge fall down and kill people and there is a frame of mind in which this obviousness is obvious enough. *How* not to have the bridge fall down is hard.

This is possibly surprising coming from the person who came up with coherent extrapolated volition, co-wrote the *Cambridge Handbook of Artificial Intelligence* article on "The Ethics of AI," etc. The relevant background comes from Eliezer's writing on the minimality principle:

> [W]hen we are building the *first* sufficiently advanced Artificial Intelligence, we are operating in an extremely dangerous context in which building a marginally more powerful AI is marginally more dangerous. The first AGI ever built should therefore execute the least dangerous plan for preventing immediately following AGIs from destroying the world six months later. Furthermore, the least dangerous plan is not the plan that seems to contain the fewest material actions that seem risky in a conventional sense, but rather the plan that requires the *least dangerous cognition* from the AGI executing it. Similarly, inside the AGI itself, if a class of thought seems dangerous but necessary to execute sometimes, we want to execute the least instances of that class of thought required to accomplish the overall task.
>
> E.g., if we think it's a dangerous kind of event for the AGI to ask "How can I achieve this end using strategies from across every possible domain?" then we might want a design where most routine operations only search for strategies within a particular domain, and events where the AI searches across all known domains are rarer and visible to the programmers. Processing a goal that can recruit subgoals across every domain would be a dangerous event, albeit a necessary one, and therefore we want to do *less* of it within the AI."

So the technical task of figuring out how to build a robust minimal AGI system that's well-aligned with its operators' intentions is very different from "AI ethics"; and the tendency to conflate those two has plausibly caused a lot of thought and attention to go into much broader (or much narrower) issues that could have more profitably gone into thinking about the alignment problem.

One part of doing the absolute bare world-saving minimum with a general-purpose reasoning system is steering clear of any strategies that require the system to do significant moral reasoning (or implement less-than-totally-airtight moral views held by its operators). Just execute the most simple and straightforward concrete sequence of actions, requiring the least dangerous varieties and quantity of AGI cognition needed for success.

Another way of putting this view is that nearly all of the effort should be going into solving the technical problem, "How would you get an AI system to do some very modest concrete action requiring extremely high levels of intelligence, such as building two strawberries that are completely identical at the cellular level, without causing anything weird or disruptive to happen?"

Where obviously it's important that the system not do anything severely unethical in the process of building its strawberries; but if your strawberry-building system requires its developers to have a full understanding of meta-ethics or value aggregation in order to be safe and effective, then you've made some kind of catastrophic design mistake and should start over with a different approach.

Norms For Link Posts

There are a lot of interesting content I read on the internet (some of which may not have been produced by the rationalist diaspora (and which the authors wouldn't want to crosspost to Lesswrong)) that I nonetheless found interesting and feel like sharing with other Lesswrongers. However, I am not sure if the posts would be well received. I have not found explicit discussion (I searched the titles in Meta) indicating what is and is not acceptable to be linked. I don't want to gain information by experimenting (which has significant negative externalities for both me and the community. Me in the form of lost karma, and the community in the form of degradation of quality, and annoyance/displeasure of community members). I would appreciate if the norms for link posts were explicitly outlined, so that I would know if a link is acceptable or not.

Points that I feel should be explicitly spelled out:

- Time: Are links older than <insert time frame here> not allowed?
- Frequency: Should we limit our link posts to a certain frequency? One a week, One a day? Not more than X a day?
- Content: What topics are not acceptable for link posts (if it is the same as for front page posts, then feel free to skip this). If link posts have a wider or narrower scope please indicate this.

I think those are the main areas for which I would appreciate explicit guidelines as far as link posts are concerned.

What Evidence Is AlphaGo Zero Re AGI Complexity?

Eliezer Yudkowsky write a post on Facebook on on Oct 17, where I replied at the time. Yesterday he reposted that here (link), minus my responses. So I've composed the following response to put here:

I have agreed that an AI-based economy could grow faster than does our economy today. The issue is how fast the abilities of one AI system might plausibly grow, relative to the abilities of the entire rest of the world at that time, across a range of tasks roughly as broad as the world economy. Could one small system really "foom" to beat the whole rest of the world?

As many have noted, while AI has often made impressive and rapid progress in specific narrow domains, it is much less clear how fast we are progressing toward human level AGI systems with scopes of expertise as broad as those of the world economy. Averaged over all domains, progress has been slow. And at past rates of progress, I have estimated that it might take centuries.

Over the history of computer science, we have developed many general tools with simple architectures and built from other general tools, tools that allow super human performance on many specific tasks scattered across a wide range of problem domains. For example, we have superhuman ways to sort lists, and linear regression allows superhuman prediction from simple general tools like matrix inversion.

Yet the existence of a limited number of such tools has so far been far from sufficient to enable anything remotely close to human level AGI. Alpha Go Zero is (or is built from) a new tool in this family, and its developers deserve our praise and gratitude. And we can expect more such tools to be found in the future. But I am skeptical that it is the last such tool we will need, or even remotely close to the last such tool.

For specific simple tools with simple architectures, architecture can matter a lot. But our robust experience with software has been that even when we have access to many simple and powerful tools, we solve most problems via complex combinations of simple tools. Combinations so complex, in fact, that our main issue is usually managing the complexity, rather than including the right few tools. In those complex systems, architecture matters a lot less than does lots of complex detail. That is what I meant by suggesting that architecture isn't the key to AGI.

You might claim that once we have enough good simple tools, complexity will no longer be required. With enough simple tools (and some data to crunch), a few simple and relatively obvious combinations of those tools will be sufficient to perform most all tasks in the world economy at a human level. And thus the first team to find the last simple general tool needed might "foom" via having an enormous advantage over the entire rest of the world put together. At least if that one last tool were powerful enough. I disagree with this claim, but I agree that neither view can be easily and clearly proven wrong.

Even so, I don't see how finding one more simple general tool can be much evidence one way or another. I never meant to imply that we had found all the simple general tools we would ever find. I instead suggest that simple general tools just won't be enough, and thus finding the "last" tool required also won't let its team foom.

The best evidence regarding the need for complexity in strong broad systems is the actual complexity observed in such systems. The human brain is arguably such a system, and when we have artificial systems of this sort they will also offer more evidence. Until then one might try to collect evidence about the distribution of complexity across our strongest broadest systems, even when such systems are far below the AGI level. But pointing out that one particular capable system happens to use mainly one simple tool, well that by itself can’t offer much evidence one way or another.

Prosocial manipulation

There is an axis of social calculativeness: whether your speech and social actions were carefully designed for particular outcomes, versus being instinctive responses to the situation.

This is related to an axis of honesty: whether your words represent your actual state. I suppose because the words most likely to produce the best response naively are often not true. Though I'm not sure if this is reliably true: feelings in the moment are often misleading, and honesty is often prudent.

Another axis is selfishness versus pro-socialness: whether your actions are meant to produce good outcomes for you (potentially at the expense of others) or a larger group such as the world.

The calculativeness axis seems widely expected match the selfishness axis well. Manipulative people are bad. I don't see why they should go together though, in theory. You can say what you feel like in conversation, or say things calculated to achieve goals. Shouldn't people saying things to achieve goals do so for all kinds of goals, many venerable? In about the same distribution as people doing other things to achieve goals?

A natural question is whether calculated behavior really is reliably selfish, or whether people just feel like it is for some reason. I can think of cases where it isn't selfish. For instance, a diplomat trying to arrange peace is probably choosing their words very carefully, and with regard to consequences. But it is hard to say how rare those are.

Perhaps we just don't think of that as being calculative? Or I wonder if we do, and while we like it if peace is arranged, we would still be somewhat wary of a very good diplomat in our own dealings with them. Because even if they are acting for the good of the world, we suspect that it won't be for *our* good, if we are the one being calculated about.

After all, we are presumably being led away from whatever our default choice would have been after hearing the person just represent their internal state as came naturally. And moving away from that sounds probably worse, so more likely that manipulation means to exploit us somehow than to secretly help us get an even better outcome. This is closely related to the honesty axis, and would mean 'manipulative' doesn't really imply 'globally consequentially bad' so much as 'dangerous to deal with'.

I am speculating. Are there common positive connotation terms for 'socially manipulative' or 'calculating'? Is that a thing people do?

Winning is for Losers

This post originally appeared on Ribbonfarm. It was written as part of the Ribbonfarm long-form *writing course and edited by Joseph Kelly. I owe Joseph and the Ribbonfarm editors (Venkatesh Rao and Sarah Perry) huge thanks for spending the time to make me a better writer.*

> Our world is filled with competition, frenzied ambition in every domain. In Western nations, and above all in the United States, it animates not only economic and financial life, but scientific research and intellectual life as well. Despite the tension and the unrest it brings, these nations are inclined on the whole to congratulate themselves for having embraced the spirit of competition, for its positive effects are considerable.
>
> *— Rene Girard, The One by Whom Scandal Comes*

I. Eating Dogs

Human life is all about competition, from the micro level to the macro.

We are built by genes that outcompeted their rivals over aeons of natural selection.

Children cooperate less and compete more as they grow older, even when competition is irrational. By the time boys and girls hit puberty they start mercilessly fighting for status, in addition to competing for resources and attention. As people enter the world of dating and finding mates, the competition for status only intensifies. With dating having moved online, everyone competes for the attention of their beloved against thousands of other Tinder matches. And sometimes also with the 5 other people they set up a date with in the same bar. The winner takes it all, and nice guys finish last.

We like exercise, music, and cooking. We like professional sports, American Idol, and cooking competitions even more.

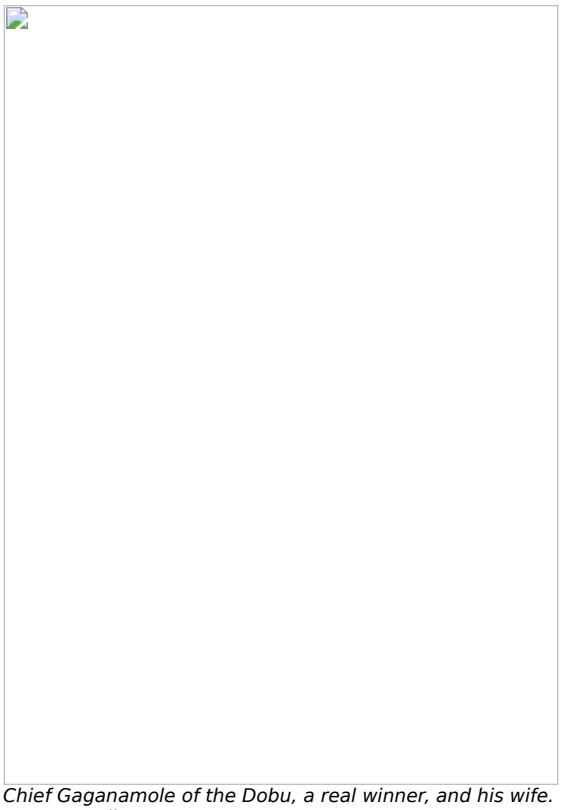
Politics is war. The political right sees a war between barbarous foreigners and a civilized America. The left sees a war between economic classes, or among a multitude of identity groups fighting to oppress each other. The Libertarian Party is the only one that doesn't look at politics as being primarily about fighting someone and they consistently gain less than 1% of the vote, the losers.

Our education system emphasizes competitive admissions, exams, and grading on a curve. This is done to prepare students to compete in the job market and the economy.

Our economy is based on companies competing with each other in the marketplace. But if you think that employees in the same company will cooperate for the good of the organization then you haven't been paying attention Ribbonfarm: organizations merely set the stage for a Darwinian contest in which sociopaths possessing the will to win oppress the clueless and exploit the losers.

If you don't spend your time thinking of ways to exploit people you're probably a loser too. You should wake up to the reality of life as a competition and follow the example from the Dobu Islanders of Papua New Guinea, a society that embraced this idea completely and without reservations. Ruth Benedict and Sam Harris describe their culture, the epitome of taking this philosophy to the extreme:

> Life in Dobu fosters extreme forms of animosity and malignancy which most societies have minimized by their institutions. All existence appears to [the Dobuan] as a cut-throat struggle in which deadly antagonists are pitted against one another in contest for each one of the goods of life. Suspicion and cruelty are his trusted weapons in the strife and he gives no mercy, asks for none.
>
> The Dobu appear to have been as blind to the possibility of true cooperation as they were to the truths of modern science. Every Dobuan's primary interest was to cast spells on other members of the tribe in an effort to sicken or kill them and in the hopes of magically appropriating their crops.
>
> […]
>
> To make matters worse, the Dobu imagined that good fortune conformed to a rigid law of thermodynamics: if one man succeeded in growing more yams than his neighbor, his surplus crop must have been pilfered through sorcery. […] The power of sorcery was believed to grow in proportion to one's intimacy with the intended victim. This belief gave every Dobuan an incandescent mistrust of all others, which burned brightest on those closest. Therefore, if a man fell seriously ill or died, his misfortune was immediately blamed on his wife, and vice versa. The picture is of a society completely in thrall to antisocial delusions.
>
> *— Sam Harris, The Moral Landscape*



Chief Gaganamole of the Dobu, a real winner, and his wife.
Image Credit: George Brown.

Huh, this doesn't actually sound so great.

The problem with living in a dog-eat-dog world is that dogs just aren't very tasty. But is it avoidable? Can you do well in life without trying to compete, dominate, and win anything? Can you even get a date?

I think so. Instead of eating dogs, we can try to bake pies instead.

II. Baking Pies

> Life is a game, play to win.
>
> — This guy, or this guy, or maybe this guy.

I disagree with all three guys, but only with the second part of the statement. Life *is* a game, but there's more to playing games than trying to beat someone. To understand this game better we require some general theory of games. I suggest game theory.

Game theory distinguishes between **zero-sum games** which are purely adversarial and **positive-sum games** which allow for cooperation. "Zero-sum" means that any gain for one player means a loss for the other players. In a zero-sum game there are no win-win possibilities and thus no point in trying to cooperate.

Imagine someone emptying a bucket full of coins (for my tech savvy readers: an ICO of cryptocoins) over a busy street. Every person in the area now finds themselves engaged in the hilarious game of looking for quarters. All the players end up with a positive outcome in monetary terms (if we ignore dignity), but the game is purely zero-sum because each coin picked up by Mr. Black is one less coin available for Ms. White to find.

If we desire to live less like the Dobu we should learn to recognize zero-sum games and avoid them. The coin game gives us two heuristics for doing that. The first is that zero-sum games usually take the form of **dividing a fixed pie**. In our example, the "pie" was the bucket of coins dumped on the street. The players have no way to get *more* coins thrown at them, they can only compete for the coins that are already there. The second heuristic is that **each player is unhappy when more and better players join** the game. As more talented coin scavengers join, fewer coins are left for you.

In contrast, a positive-sum game involves a collaborative effort to which many players can contribute. Players **bake a bigger pie by cooperating**. In positive-sum games, the entrance of new participants is either bad or good for the incumbents, depending on the situation.

Let's look at a more complex example from an arena that at first glance appears purely competitive – professional sports. Specifically, is the NBA a zero-sum or positive-sum game for LeBron James?

LeBron James competing. Image credit: Ezra Shaw / Getty Images.

A single game of basketball is a relatively zero-sum affair, but athletes don't join the NBA for the pursuit of basketball wins in a vacuum. They get many rewards for participating: money, fame, groupies, and the satisfaction of a basketball game played at the highest level. All of those make up the pie that NBA players bake together.

The title of "NBA Champion" is a yearly zero-sum game, but it's an artificial format invented by the league. If the league could sell more tickets by having multiple concurrent champions or by awarding style points instead of titles, it would.

LeBron welcomes better players joining the league because that would increase the NBA's prestige, popularity, and profits, of which he gets a share. In fact, in 2017 LeBron cost himself money by beating other teams *too quickly* – this led to fewer playoff games, which in turn decreased league revenues, total salaries paid to players, and subsequently the value of LeBron's own contract. LeBron wants the league to be as good as possible, and the other players are collaborators rather than competitors in the bigger picture game of the NBA.

Of course, the NBA looks much more zero-sum to a marginal player. Unlike LeBron, a benchwarmer is *not* happy when more talent joins the league, they may end up taking his job. This points to another important principle of games: **strong players have more room to cooperate, while weaker players are forced to compete with each other**.

Let's consider education, specifically going to a prestigious university. If you're a borderline candidate for a university, strong applicants reduce your chance of admissions. Once you're in, they make your grading curve steeper and compete for on-campus leadership positions and ultimately for jobs. Competing against stronger students can have demoralizing effects that persist long after school is over.

This isn't the case for the student who is much smarter than her peers. She *welcomes* stronger classmates. They improve her learning opportunities and increase the overall prestige of the university, without being a threat.

There are two ways to become a stronger player and "rise above the competition," as it were. You can try to outwork everyone else, or you can look to be a bigger fish in a smaller pond. Both options can work for a college applicant, although probably not as much for a basketball player. The NBA is the only game in town, and NBA players are presumably already working as hard as they can.

However, there's another way to avoid the grind of competition: instead of being the **strongest** player, be the **strangest**.

If you possess a unique skill, it complements the skills of other players instead of competing with them. NBA players have built lucrative careers as "the guy who just blocks shots and has a sweet 'fro" or "the white guy who just stands in the corner and makes threes." It's enough to do only one thing well if that thing is rare.

But for avoiding competition, having unique skills isn't half as important as having unique *desires*. The philosopher René Girard described the mimetic contagion of desire: people instinctively imitate the desires of those around them, which leads to everyone chasing the same prizes. These prizes often have no inherent value other than being the objects of shared pursuit. When those prizes are in limited supply, this pursuit creates zero-sum competition and leads to bitter rivalries.

The two ways of being similar reinforce each other. When people go after the same prizes, they will develop similar skills in the pursuit. When people's skills don't set them apart, the will try to stand out by competing ever more desperately for the common prizes.

We talked before about how prestigious universities set the scene for endless competition among the students at every stage of their education. Dan Wang ties this to Girard's idea of competition stemming from similarity and mimetics:

> The closer we are to other people—Girard means this in multiple dimensions—the more intensely that mimetic contagion will spread. Alternatively, competition is fiercer the more that competitors resemble each other. When we're not so different from people around us, it's irresistible to become obsessed about beating others. […]
>
> It's hard to construct a more perfect incubator for mimetic contagion than the American college campus. Most 18-year-olds are not super differentiated from each other. By construction, whatever distinctions any does have are usually earned through brutal, zero-sum competitions. These tournament-type distinctions include: SAT scores at or near perfection; being a top player on a sports team; gaining master status from chess matches; playing first instrument in state orchestra; earning high rankings in Math Olympiad; and so on, culminating in gaining admission to a particular college.

> Once people enter college, they get socialized into group environments that usually continue to operate in zero-sum competitive dynamics. These include orchestras and sport teams; fraternities and sororities; and many types of clubs. The biggest source of mimetic pressures are the classes. Everyone starts out by taking the same intro classes; those seeking distinction throw themselves into the hardest classes, or seek tutelage from star professors, and try to earn the highest grades. […]
>
> No one has ever asked me how one should escape mimetic contagion on campus. Still here's my answer: If one must go to college, I advise cultivating smaller social circles. Instead of going to class and preparing for exams, to go to the library and just read. Finally, not to join a fraternity or finance club, but to be part of a knitting circle or hiking group instead.
>
> — Dan Wang, College as an Incubator of Girardian Terror

Most of the prizes students compete for aren't really worthwhile even when the temptation to compete for them is overwhelming. Is the point of attending college to be elected finance VP of some fraternity? College should be a place to have fun, get laid, make friends, learn something, and figure out which career suits your individual skills and tastes. These are mostly cooperative pursuits, and Girardian competition stands in the way of achieving them.

The **strongest** and **strangest** (e.g. knitting circle) students won't get sucked into the competitive vortex. They'll spend time in the library studying whatever weird subject they're obsessed with, they'll make friends with fellow geeks, and they'll wonder why most of their classmates are perpetually miserable.

III. Tits and Tats

We have started building a framework of competitive and cooperative situations. Competition stems from zero-sum contests over a fixed pie, where additional players are never welcome. Cooperation comes from an opportunity to bake a pie collaboratively, and strong players are welcome if they contribute. Ending up in the latter situation requires being more capable than anyone else, or really different from everyone else.

This foundation is enough to survive in Harvard or the NBA, but it's insufficient for a real challenge like OkCupid. For a strategy that works in online dating we need to dig deeper into game theory, and the one particular game that is most heavily theorized about.

The same way that biologists are supposed to study *all* living creatures but end up mostly focusing on mice and fruit flies, so it is with game theorists. They're ostensibly studying *all* possible games, but a huge chunk of the literature is dedicated to a single one, the prisoner's dilemma. There are many analogous framings of the dilemma, I prefer this simple, prisoner-free formulation:

You receive a widget with two buttons on it, labeled **"cooperate"** and **"defect."**You are informed that another person somewhere in the world received the same widget. If you press "defect," $1,000 will be immediately deposited into your bank account while the other player, whom you'll never meet, gets nothing. If you press "cooperate," the other person gets $3,000, but *you* get nothing except for a warm feeling. You both make your choices without knowing what the other person chose.

That's it, that's the game.

The salient feature of the prisoner's dilemma is that choosing "defect" makes a player $1,000 richer *regardless of what the other player is doing*. In the absence of mechanisms to influence each other, this usually leads to both players defecting. Of course, if both players chose to cooperate they'd each be better off by $2,000.

The "strong or strange" principle applies here as well. A billionaire may be happy to let a random person take $3,000, so may the guy who lives out of a van and climbs giant cliffs without a rope. The former has enough money and the latter doesn't even need it to buy a rope. But for people who are neither very strong or very strange cooperation is difficult and defection is tempting.

This simple setup belies a rich universe of human interaction. In his book, Moral Tribes, psychologist Joshua Greene shows that most of our social intuitions and moral emotions evolved as means to cooperate in prisoner's dilemmas with other people. *Empathy* and *compassion* allow players to cooperate by making it intrinsically rewarding to benefit others. The capacity to feel *self-righteous* or *guilty* signals a personal commitment to doing the right thing. Emotions like *tribalism* and *loyalty* allow cooperation to be enforced by a broader collective or by an authority figure.

The easiest way to achieve mutual cooperation is by repeated prisoner's dilemmas played with the same partner. This allows each player to play tit for tat – reward a cooperator with cooperation in the next round of play, and defect against a defector. *Tit for tat* is implemented in nature by everyone from fish to birds to monkeys. It's such a useful cooperation strategy for Homo sapiens that we evolved a whole suite of emotions that help us implement it: *anger, trust, vengefulness,* and *gratitude*.

Tit for tat works when you're dealing with the same **few players** over **long time-frames**. The strategy doesn't work if players don't expect to interact in the future. The incentives of future cooperation or punishment lose their bite when dealing with large groups of players, or when those players are only concerned with immediate outcomes.

Tits on a tat

What kind of game is online dating? It can be a short-term and multi-player game in which everyone screws each other (in the literal and good sense, but also in the figurative and bad sense). But if two people are trying to build a real relationship, dating needs to be a cooperative, long-term, two-player game. If that's the game you're playing, *tit for tat* is your strategy.

Now, it may seem obvious that finding a romantic partner should be a collaborative pursuit and not a hostile contest, but it's not the *natural* approach. The way of dating in nature is spiky dicks.

IV. Spiky Dicks

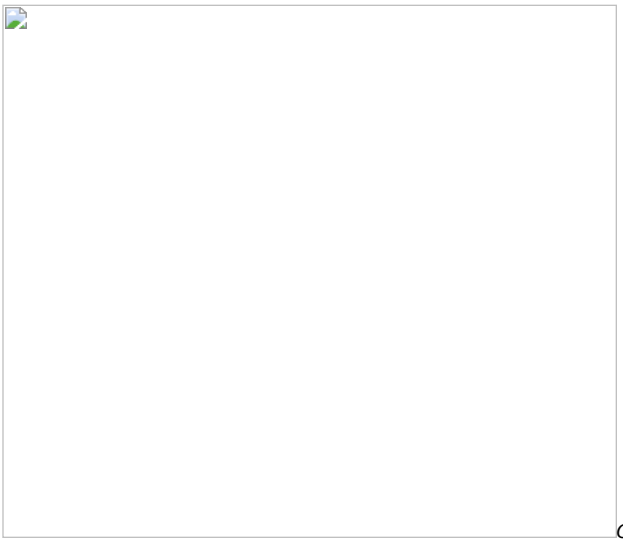
> "Everything in the world is about sex except sex. Sex is about power."
>
> — Oscar Wilde

Wilde's famous quote summarizes all the research in evolutionary biology which shows that throughout the animal world sex is about competition, and competition is usually about sex. Animals may cooperate with each other to acquire food and avoid being eaten. But as soon as that's taken care of, it's back to vicious contests over mating.

We normally think of mating-related competition as happening among members of the same sex, particularly males. For example, male elephant seals fight so savagely for access to females that at the end of their mating season 4% of males will have had most of the sex but 90% of males will carry scars and injuries from fighting. From Achilles vs. Paris to Swaggy P vs. D'Lo, many a historic beef among men has started over a woman.

But the real action is in male vs. female conflicts. Those are no less violent, and often a lot more creative.

In several species of beetles, the males have evolved sharp spikes on their genitalia which anchor the female in place during copulation. Female beetles evolve more soft tissue in the copulatory duct to protect themselves from injury, which in turn leads males to evolve ever scarier looking dickheads.



Callosobruchus analis penis (beetle dick). Image credit: Wikipedia.

Ducks have taken this idea one step further, and then fifty more steps in really weird directions.

Instead of chocolate and roses, male ducks usually go for the "forced copulation" approach to dating. In response, female ducks evolved corkscrew vaginas, which made male ducks evolve spring-loaded foot-long corkscrew penises (with spikes on them, of course). Finally, female ducks evolved branching labyrinthine vaginas so they can send the sperm of a male they don't like towards a literal and reproductive dead end.

Seriously.

This sort of sexual arms race is the norm in the animal world. So far we humans haven't sprouted spiky genitalia, our main weapon in inter-sex conflict is lying and deception. Members of both sexes pretend to be fitter and more faithful to their partner than they really are. Better pretense leads to better detection of trickery, which leads to ever more sophisticated lying. Eventually, people evolved the ability to convincingly lie to themselves, all the better to fool others about their commitment and attraction to a potential mate.

This is as true today as it was on the savannah. In online dating men lie about their height and income, women lie about their age and weight, and a quarter of profiles have photoshopped pictures. And you thought the news was fake.

Bullshitting is a useful strategy for spreading your genes widely with a minimal commitment of resources, or for beating your roommates in a competition to sleep with more women. Most lies don't survive beyond the first date, but they get a lot of people to go on that first date and get drunk enough to jump into bed with you. This is a very **short-term** and **multi-player** approach to dating, and some people assume that this is the only one.

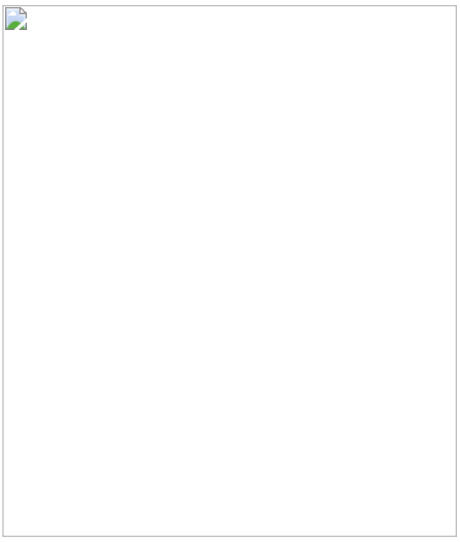
Everyone complains that while online dating made it easier to get a first date, turning that first date into a relationship became a lot harder. Most dating advice answers this conundrum with "keep doing what got you the first date, just more and better." This is dumb and doesn't work. What gets you first dates is mass-appeal and lying. Those are defection strategies, they benefit the player while making dating harder for both the gender they pursue and the one they compete with. Online dating isn't defective, it's the players who keep defecting.

Getting one person to spend a thousand nights with you is the exact opposite of getting a thousand people to spend one. It requires playing the opposite kind of game: **long-term** and focused on a **single person** (or three, but not fifty). The strategy in this sort of game is to play *tit for tat* to achieve mutual cooperation with the person you will eventually end up with. You can play cooperatively with that person *even if you haven't met them yet.* In fact, your first shared goal is to find each other, and then build the foundation for a relationship that will make both of you happy.

The first step towards this is complete honesty. If the other person is so ambivalent about meeting you that an inch of height or a year of age would tip the balance, you probably won't end up picking baby names together anyway. You shouldn't lie on your profile even if everyone else does. The novelty of seeing someone who fulfills exactly what their profile promised will kick off your first dates on a note of pleasant surprise instead of disappointment.

But just being honest is not enough. In accordance with "stronger or stranger," to avoid competing with everyone else for your partner's attention, you have to be really irresistible or really weird. The latter is much easier and works just as well as the former.

OkCupid's data shows that conventionally attractive profile pictures get far fewer messages than photos that elicit strong positive and negative reactions. As long as at least a few people really dig you, having a lot of haters is not to your detriment. When I started dating online, I wasn't sure if I should use the photo below as my main profile pic. But when two women wrote me just to say that they would never date a man with a photo like this, I knew this was the right one.



My goal was to make it easy for my as yet unknown partner to find me, so I made my profile idiosyncratic enough to filter out most of the users that weren't *her*. Instead of a self-summary, I started the profile with a stupid poem. I mentioned all my esoteric interests like Bayesian epistemology. I listed several reasons *not* to date me. As I kept making my profile quirkier, the women it attracted were a lot more interesting to me.

Finally, I got my reward:

The point of *tit for tat* is to defect against defectors (the 99% of women who aren't really into me) and to cooperate with cooperators (the few who are). For profile design, this means scaring away the people who are attracted to you superficially and appealing to those who like your unique quirks. On first dates, it means cutting off those who aren't ready to risk making a small commitment to you, and building something with those that are.

At the start of a relationship, the "defect" move is to go along on a few dates while swiping for other matches in the meantime. It keeps your own options open but does the opposite for the person you're seeing. The temptation to do this exists because the 1,000 potential people you haven't met yet appear perfect in the fuzzy light of imagination, while the actual person in front of you has shown a wart or two. But mutual defection has costs: it prevents both partners from making the effort to build the relationship on a stronger foundation than just mutual lust. Without that foundation, the lust hormones dissipate after a couple of months and both people are back where they started, slightly frustrated and two months older.

Tit-for-tatting the first date mostly means going against common advice.

Everyone says to avoid heavy topics on the first date. But why would you waste time with someone with whom you can't have a serious conversation about the meaning of life or the minimum wage? If these topics aren't deal breakers, you should be able to talk about them with open-mindedness and humility. If they are, you should use them to filter out incompatible matches and get back to looking for the ones who understand labor economics.

Everyone says to avoid talking about your ex on the first date. Maybe that's a good idea, but for an entire year while I was dating, I shared a one bedroom apartment with an ex-girlfriend. In New York City, it takes more than a broken heart to give up paying half-rent for a sweet pad. When I started dating again I didn't feel comfortable bringing this up, but then I realized that I should talk about it unapologetically on the first date. I would ask my dates to trust me that this was a temporary habitation circumstance, not a permanent emotional one.

This confession actually worked to my advantage, it sent a strong signal that I have nothing to hide and there were no other shoes waiting to drop. The women who were willing to trust me reciprocated by telling me something embarrassing about themselves, and we turned an awkward situation into an opportunity to build mutual trust.

Everyone says to hold off on texting after the date, so as not to appear desperate. I assumed that if I had an honest and deep conversation with someone on the first date, she would have plenty of information about my value as a romantic partner without having to deduce it from the timing of my text. Waiting 3 days to text creates uncertainty, and in the cooperative game of dating uncertainty increases the odds of defection.

Instead of the 3 day rule, I went with the -1 day rule. If I enjoyed the date I would say: "Hey, I really enjoyed this date! I'm going to text you tomorrow at 8 pm to see if you want to go on another one." This is a *tit for tat* move. I'm clearly showing that I'm playing *cooperate*, and I set clear expectations for reciprocation. Because I let the girl know in advance when I'll reach out, if I don't get a reply relatively quickly the next day I can safely assume that she's not interested, which saves me from chasing ghosts.

One piece of common wisdom that *is* actually true is that vulnerability is the key to building intimacy. And yet, very few people are willing to be vulnerable in front of potential romantic partners. Vulnerability is the ultimate *tit for tat* strategy: there's a lot to gain if the other player reciprocates, and a lot of pain if they defect.

Of course, it's possible to be *too vulnerable* on the first few dates, just as it's possible to be too weird, too deep, too honest or too demanding. But in my experience, people are afraid of being too open much more often than they actually are.

The strategies above *do* often fail, in the sense of scaring someone away from a second date. But if they only "fail" in cases where the second date wasn't going to lead to a tenth, that's a feature. As with startups, if you're going to fail you should fail quickly, and move on to someone who tits your tats.

And when those strategies succeed, they do so magnificently. If you lead off the first date with honesty, vulnerability, and commitment and the date turns into a relationship, the relationship will also be based on honesty, vulnerability and commitment. This is worth a lot when so many relationships are based instead on pretense and power games.

Tit for tat doesn't apply to all aspects of a first date. Where it doesn't, just cooperate unconditionally. Take a shower, show up early, commute to the other person's neighborhood, turn off your phone, offer to pay. Don't be a spiky dick.

I learned those lessons over a couple of years a couple dozen OkCupid dates. Like every game, dating is a skill that improves with practice. With my *tit for tat* game and with the help of a spreadsheet, Bayesian epistemology girl and I ~~are getting married in the fall~~ *hopefully got married in the time between me writing this post and publishing it. If we didn't, it's going to be really awkward*.

V. Fighting Moloch

So far I've talked about how cooperating instead of trying to beat someone leads to personally beneficial outcomes. But there's more at stake here than your next first date.

Mutual cooperation gets harder the more players are involved. At the extreme, a prisoner's dilemma played by an entire society often results in everyone defecting against each other. As in the two person game, "defect" is any move by a player that nets them a small gain while imposing a large cost on others. Here are some examples of defections in society-wide games: sending a marketing email, using antibiotics, burning some coal, calling someone a Nazi online. The corresponding outcomes: pervasive spam, drug resistance, global warming, Twitter. These outcomes are common and tragic, so they're known as tragedies of the commons.

There's a view that failure to cooperate on multiplayer prisoner's dilemmas is the greatest threat to our civilization, or *any* civilization for that matter. This position is best articulated by Scott Alexander, who gave it a name: Moloch.

Moloch is why, when food is scarce, the animals (and humans) that breed and kill most efficiently outcompete and destroy those that don't. *Moloch* is why governments race to the bottom and provide corporate welfare. *Moloch* is the force behind arms races, environmental destruction, and clickbait – competitions that leave every single participant worse off.

Humanity currently enjoys a moment where the resources available to us exceed our ability to exploit them. We can afford to engage in activities that aren't part of a ruthless competition for resources: art, leisure, blogging. But once our capacity for exploitation increases – for example with the advent of smarter-than-human AI – art, leisure and blogging will become unaffordable luxuries.

Scott offers an escape: transhumanism. The goal is to create something or someone that shares our values, nd is so strong that it doesn't have to sacrifice those values for the sake of competition.

I know, I know, this sounds pretty insane. Whether one thinks that this plan is feasible or not depends on many things, like one's geographic distance from the Bay Area. But here's the fun part – it's a great way to fight Moloch even if it doesn't work.

Imagine if we were trying to design a community of people devoted to cooperation, based on everything we learned about competitive and cooperative games. How should we approach this?

We would build a community dedicated to creating something new, a freshly baked pie. It would have to be a long-term project. It would have an important and purely collective reward at stake, like protecting against a common tragedy. It would involve a bunch of weirdos.

It would be something like transhumanism.

Transhumanism inherently creates a cooperative culture among those involved in it. The pursuit of an outlandish goal in the far future, like friendly AI, cryonics, curing aging, or hastening the singularity, is a remarkable way to turn naturally uncooperative geeks into a collective.

Does that make transhumanism sound like a religion? The two main faults of religions are that they turn their followers to violence against the outgroup, and that they untether their followers from reality. Encouraging their followers to cooperate and to think long-term is overall a positive aspect of religions. Transhumanists try to be attuned to the physical and technological reality, and the ingroup of transhumanism is the entire human species. As far as religions go, it gives you most of the good stuff with little of the bad.

Of course, it's hard to join a community you don't believe in just for the benefit of a cooperative culture. There's another way to achieve the same goal: create that culture yourself. Ultimately, "culture" is just a set of norms that people follow. You don't need a community to start living by those norms yourself, and watch them spread to those around you.

Whichever game you're playing, lead with cooperation and play tit-for-tat. Cooperate at times even when the other person seems to defect, just in case. Be honest and radically transparent to reduce the cost of interacting with you. Pursue weird interests and goals. Write honestly about your weird interests and goals, and publish them for free online. Don't be a dick. Deal with every person as if you're going to be playing repeated games with them for the next 10,000 years.

If the transhumanists get their way, it may actually happen.

Frequently Asked Questions for Central Banks Undershooting Their Inflation Target

If you are a central bank undershooting your inflation target, please read this first before posting a question about how to create more inflation.

Q. Help! I keep undershooting my 2% inflation target!

A. It sounds like you need to create more money.

Q. I tried that, and it didn't work!

A. How much money did you create?

Q. Five dollars.

A. Okay, now we know that five dollars wasn't enough money. You need to create more.

Q. More than five dollars?

A. Right.

Q. I've heard that it's very bad when governments create lots of money just so they can buy things. Prices skyrocket, and soon nobody wants to hold on to the money anymore! That's why we have stern and independent central banks, to prevent too much money from being created.

A. Yes, that's a big problem, all right! That problem is the opposite of the problem you actually have. It's definitely true that if prices start going up faster than you want, you should stop creating money and maybe destroy some of the money you already made. But that is not the problem you have right now. Your current problem is that there's too little money flowing through the economy, meaning, not enough money to drive all the buying of real goods and labor that could be exchanged if your economy had more money.

You know how hyperinflation happens because there isn't enough real stuff and so the tons of new money are just competing with other money to buy a limited amount of real stuff? One way of looking at your 2% inflation target is that you're supposed to create just enough money flow that it's a little above what's required to animate all the trades your economy can make. If there's a little more money flow than the minimum required to animate all potential trades, that money is competing just a little with other money to buy stuff. Producing, say, around 2% inflation. That's why your government gave you an inflation target that was low, but not super-low and not zero. Right now your economy doesn't have that much inflation, because there isn't enough money flowing and your economy is failing to make all the trades it could make. That's why 'undershooting your inflation target' is associated with a country that feels sad and listless, with all the factories and shops still there but people not having enough

money to buy things from them, because not enough customers are buying from their own enterprises. It's very depressing.

Q. Can I solve that problem by being a stern, independent central bank that refuses to create more money?

A. No.

Q. How about if I sit down to lunch with some big banks and ask them to make more loans, maybe accompanied by a significant look where I have one eyebrow raised?

A. Even if they listened, I'd worry that everything in your economy is currently in equilibrium and other banks might make fewer loans once there were fewer loan opportunities left. Anyway—I'm just guessing here—did you maybe already try that?

Q. Yes.

A. And what happened?

Q. It didn't work. But that doesn't mean it won't work next time!

A. Your economy, and the people in your economy, are suffering right now. You should be creating more money right away, not letting innocent people suffer while you experiment with weird alternatives to action.

Q. But I don't want to create more money!

A. Then you won't get more inflation. If you want more inflation, you need to create more money.

Q. But I already printed five whole dollars and prices didn't go up at all! They actually fell! I'm starting to wonder if printing money really makes prices go up.

A. There were probably some banks exploding at the same time you were printing the five dollars, and the exploding banks destroyed more than five dollars. Even though you're the only entity that's allowed to create your country's base money, banks also create a kind of virtual money when they make loans. This means that when a bank explodes, it can destroy some virtual money. So even though you printed five dollars, the bank exploding at the same time destroyed more than five dollars of virtual money, and the total amount of money went down. That might be why prices didn't rise. Or it might be more subtle, like people wanting to hang onto the money they already have. That also decreases the speed at which money changes hands, which means that there's less effective money per transaction in the total economy, which puts downward pressure on the price of each transaction.

Q. So there's nothing I can do? That's terrible! But I guess if there's nothing I can do, it's not my fault—

A. No, you can do something! You just need to create even more money.

Q. (*Grudging sigh.*) How much more money do you think might be enough?

A. Well, predicting that is a very difficult job! There are all sorts of things that affect prices besides the amount of base money, like bank lending rates. Now, all these other factors merely affect the amount of base money that's "enough", they don't mean that you could print infinite money without increasing prices. Ideally, you'd

create a prediction market to forecast the effects of printing different amounts of money. But if you create some money and prices don't increase, that definitely means you didn't print enough.

Q. What if I print enough money to make prices rise, and they still don't rise?

A. Then you were wrong about how much money was 'enough'.

Q. That seems unlikely. Maybe I'm already doing the right thing, and it's just going to take a while to work, so I don't need to change anything?

A. Sorry, but that's a definite no! The weak form of the efficient markets hypothesis says you can't have publicly predictable price changes in a liquid market. If the price of your currency was predictably going to drop later, people would short-sell it now, or just refuse to buy it at a price that would predictably go lower. So if what you're currently doing, and any future actions you've already announced, aren't already making prices higher, we already know it wasn't enough. You can also check and see if the market is pricing inflation-adjusted securities in a way that shows the market expects inflation over the next few years. If they don't, you're doing something wrong.

Q. But my expert economists say that the people who price inflation-adjusted assets are wrong, and there will be lots of inflation next year! In fact, I'm already starting to think about raising rates now to prevent that, like a stern and independent central bank should.

A. Is it possibly the case that your in-house experts have been wrong every single time they predicted more inflation than the market forecast, over the last fifteen years or so?

Q. Yes, but that doesn't mean they'll be wrong this year!

A. I'm sorry, but it sounds to me like your experts just don't have the incentives to make correct forecasts. Or maybe they lack the sheer knowledge and computational power to make better forecasts than the hedge-fund managers who can make billions and billions and billions of dollars if they predict 1% better than the market. Although… I have to say, your in-house experts being wrong in the same direction every year doesn't sound quite so innocent.

Q. But… my in-house experts have expensive suits! And credentials! Hedge-fund managers don't have suits that nice, I bet.

A. Actually, they kind of do. More importantly, they're paid literally billions of dollars to get the answer right. Maybe your in-house experts are telling you what they think you want to hear. Maybe they're just making the same innocent mistake repeatedly. But if you've already seen your in-house experts be wrong lots of times before, and they're mistaken in the same direction every time, then you need to stop listening to your in-house experts when they predict lots of happy inflation. You should pay attention to the market forecasts instead, because highly liquid market prices almost never change in a predictable net direction. When liquid prices change in a predictable net direction, it corresponds to free money! Lots of highly intelligent organisms in your financial ecology really like to eat free energy, and when they consume the free energy it eliminates the directional error. Which usually means liquid markets aren't repeatedly wrong in the same direction, like your in-house experts are. Right now, the market forecast is telling you that you need to create more money if you want inflation.

Q. It does seem clear that I should lower rates to nearly zero at my meeting next month. That's creating more money, right?

A. Sorry, let me rephrase. The market has already guessed that you plan to lower rates at your meeting next month. If the market rate for inflation-adjusted securities doesn't already imply that they expect inflation, it means they already don't expect you to create enough money. If the market doesn't already expect inflation, you need to create more money than they're currently expecting you to create. If you're worried about a self-referential circularity if you did start paying attention to the market forecast and the market realized that, you can just create a separate prediction market to directly forecast the amount of money you'll need. But right now, when you're not paying attention to the market forecast, the situation is clear—the market expects your current behavior and comfortable habits to fail, and you need to do more.

Q. But I can't lower interest rates below zero!

A. First of all, yes you can, several countries are trying it and nothing bad is happening to them. And second, you can always create money. Create new ones and zeroes and inject them into the economy. Never mind thinking about interest rates. If you create enough money, prices will go up.

Q. What if I say I'll print ten dollars and the market still thinks that's not enough?

A. Create even more money. Look, imagine creating a quadrillion dollars. Prices would go up then, right? I mean, a 12-year-old raised by goldbugs could understand that part… uh, it's possible you might need to add a 12-year-old raised by goldbugs to your advisory staff.

Q. Just because creating enough money would make prices rise, doesn't logically imply that if prices don't rise then I haven't created enough money!

A. Actually—

Q. Really honestly, I already created a large amount of base money! I'm not lying, I really made a lot! Doesn't that mean I'm already being super-loose with my policy? I just can't understand how, with my base money supply at a level of over nine dollars, you think my monetary policy is too tight.

A. The absolute amount of base money has no meaning apart from monetary velocity. 1 trillion base dollars and a bank-lending multiplier of 33 is more effective money than 6 trillion base dollars and a bank-lending multipler of 4. That's why central banking isn't as simple as just increasing the base money supply by 2% per year, or something like that. In fact, there are positive feedback cycles which means that targeting a base money level can produce wild instability. When money is becoming more valuable, people try to hold onto it more, which slows down velocity, which decreases the effective amount of money available per transaction, which decreases prices even more, which makes money even more valuable. Which increases the real burden of debt, which means that fewer people pay back their loans successfully, which blows up banks and makes them more reluctant to lend, which futher decreases the money supply, which further decreases the money available to pay back debt. So as a central bank, you need to keep your eye on the amount of money being spent and moving around, not the amount of base money that exists. If the amount of money spent and moving around is going down, or just increasing slower than it used to, you're in

trouble. And you need to do something right away, because of positive feedback cycles!

Q. I am doing something! My interest rates are super-low right now. People can take out loans super cheaply! Doesn't that mean I'm already being super-loose with my policy in a way that's just bound to create lots of inflation starting, you know, any minute now?

A. Nope! Think about Freedonia, which is printing too much money and has 100% per year inflation. Would ten percent interest rates be 'tight money' there?

Q. Ten percent interest? That's super-high!

A. Not in Freedonia! In Freedonia, if you have a company that's growing five percent real growth every year, plus one hundred percent inflation, that corresponds to 105% nominal growth. In a country like that, if you make a loan at ten percent nominal interest, it's −90% real interest! Which is 95% below the rough vicinity of where we might find the Wicksellian interest rate! And that's super-inflationary! Conversely, in Japan where there's near-zero inflation and lots of saving and an aging population, the Wicksellian equilibrium interest rate is negative-something percent, so a nominal 1% percent interest rate might be a high rate of real interest that made for tight money and produced more deflation. That's why central banking isn't as easy as clamping the nominal interest rate at three percent and holding it there forever. In fact, if you clamp nominal interest at 3%, your currency will inevitably blow up into deflation or hyperinflation! Lower inflation makes a fixed nominal rate be tighter money which produces an even lower price level. Higher inflation makes a fixed nominal rate be a lower real rate which corresponds to even cheaper money. Nominal interest-rate targeting as a control instrument is kind of silly to begin with, honestly! It's like trying to steer a car with a wobbling, unsteady steering wheel, where the same steering-wheel position might be pointing the road-wheels in a different direction every time the car goes another meter.

Q. All this sounds very complicated. Can I take a long time to think about what's going on, and maybe respond very timidly and weakly?

A. No! When your currency is gaining value, it makes people more reluctant to spend the currency, which decreases the effectively available supply of the currency, which increases the price of the currency. When banks blow up and vaporize virtual money, or when banks become more reluctant to make loans, it decreases the money flow available to pay off all the loans in the system. When inflation goes lower, it increases the real interest rate represented by the nominal rate you target, which is tighter money, which puts further downward pressure on prices. These are positive feedback loops! You need to respond right away, before the positive feedback gets out of control. If you do nothing, the loops will blow up. If you do something, but too little, they'll blow up slower. You need to do enough to interrupt the positive feedback cycle!

Q. Like sharply raising interest rates to combat inflation, before people start trying to get rid of the money they're holding and it turns into hyperinflation?

A. Right! And if inflation was still spiraling out of control, you'd raise interest rates further and, more importantly, create less money or even destroy some money! You'd have to do that right away before things got even worse!

Q. Exactly! I'd raise rates as soon as I saw inflation coming, I wouldn't wait for it to happen. I'd be super-proactive!

A. Well, you'd raise rates if the market said too much inflation was coming. You wouldn't listen to your in-house experts who've been wrong in the same direction every single time.

Q. I wouldn't?

A. Anyway, you know how you need to be super proactive and alert to prevent too much inflation? This is the same situation, only in reverse.

Q. But… it's just…

A. What is it?

Q. As a central bank raised in the modern era, I just feel deeply bad inside about inflation, you know? Even if I have to do it sometimes, it feels dirty.

A. Yes, I've gathered that.

Q. I mean, back in the day, there were bad people who created a lot of new money, and heroes who stopped them. Even though nobody believed in them, even though the world scorned them, even though lots of people were claiming that inflation had nothing to do with the money supply and was caused by labor unions or something, they still tightened monetary policy! I want to look in the mirror and see a hero, not a villain. Like the days of yore when Paul Volcker rode into battle against inflation with President Carter by his side like a loyal shieldbearer, and Carter nobly sacrificed himself so Volcker could go on…

A. Those heroes were fighting a problem that is the opposite of your problem. They correctly did the opposite of the thing you need to do. Or, on a meta-level, they did the meta-thing you need to do—they showed courage and conviction, pointed in the right direction, even though some people were claiming that the central bank was powerless to help.

Q. But what if I'm too courageous and then I overshoot my inflation target?

A. Both lab experimentation and macroeconomic history shows that price-setters are much more reluctant to lower prices than raise prices. Employers and employees are much more reluctant to lower wages than to raise wages. Most employers would rather fire one person than try to negotiate 5% salary cuts with 20 people who would all be demoralized. On the other hand, if there's enough nominal money coming in and the wage market is heating up, they're often okay with giving everyone 5% raises. This means that making your money policy slightly too tight does much more damage than making it slightly too loose, because it's a well-tested empirical fact that the people inside the economy have a much easier time adjusting to slightly higher money flow than slightly lower money flow.

Big deflation and big inflation both do enormous amounts of damage, and hyperinflation is worse because it goes further and faster. But money that's a little too tight does much more damage than money that's a little too loose! So it's really important that you create more money right now, and don't worry so much about the possibility of overshooting your inflation target a little, especially when you've undershot your target a lot up until now.

Q. But with all these dangerous positive-feedback cycles… what if prices skyrocket? Just because printing a quadrillion dollars would create hyperinflation doesn't mean

that there's a smaller amount of money I can print to cause exactly 2% inflation! There could be nonlinearities in the market.

A. Indeed there are! So here's what you do. Instead of literally printing cash, create electronic money in your own account that you then use to buy something else. Buy an asset you can sell back later, like government bonds. That way, you're creating some new money now. But you keep the assets you buy with the new money. That way, you can sell back the asset later to destroy the money after velocity picks up, or if you accidentally created too much.

Q. That sounds really wobbly to me. I'm afraid something will go wrong, and that price levels will end up going back and forth or maybe out of control.

A. There are several tools you can use to prevent that! The most important tool is to target a price path. That means, instead of constantly trying to eyeball the economy and guessing in an ad-hoc way whether it has too much or too little money, you have a definite rule that says, "If inflation goes over 2%/year then I will create less money and if inflation goes under 2%/year I will create more money, and the further over or under the level it goes, the more money I'll destroy or create." And if people trust you'll do that, they won't expect inflation to go much over or under 2%/year, and they won't value money any more or less than a long-run 2%/year level implies.

Q. So every year I try for 2% inflation that year, and miss, and end up with 1% inflation instead, and then I do the same thing again next year? Sounds like a great policy to me! I'm already on the ball when it comes to that one!

A. No, no, no! If you're doing that, you're not actually targeting the path of anything! Let's say a donut costs $1 this year, in year 0. At 2% inflation, it would cost $1.02 next year, and $1.04 in year 2, and $1.22 in year 10. Let's say that you miss your inflation target for this year and the donut ends up costing $1.01 next year. If you just shrug and say, "Oh, well, I'll try for 2% again," then your new plan is for the donut to cost $1.03 in year 2, instead of $1.04. And $1.21 in year 10. As a result of getting less inflation than you wanted, you changed your monetary target to be tighter—you changed your mind and aimed to have a donut cost less in year 10, then you previously said a donut should cost in year 10. That's exactly the sort of thing that promotes positive feedback loops!

Q. What do you mean, I'm tightening my monetary policy? I'm still saying "I want 2% per year inflation", right?

A. Before, the markets expected you to target $1.22 per donut in year 10. Now they expect you to target $1.21. That's a tightened monetary policy! And what's much worse is if you do this every year, and the markets rapidly realize that in real life the donut will only be $1.10 in year 10. Now the market expects a much tighter monetary policy and they'll consider money much more valuable than you say you want it to be. And people will hold onto even more money. Which will make you 'miss' your inflation 'target' even more. Which will make you tighten future price targets again, and so on.

Q. So what should I do instead? Tell everyone that I really, really want 2% inflation next year?

A. To target the path or maintain a level target, if you get 1% inflation in year 0, you have to create even more money in year 1 to make up for the missed target in year 0. Even if year 1 comes out to $1.01, you still have to target $1.04 for year 2, $1.06 for year 3, and so on. That way, it doesn't matter as much if you overshoot in one year,

because the stance of long-term monetary policy won't have changed in response to that. The same theory also works for combating hyperinflation—you just create less money next year, or destroy money, if you overshoot. And if you stick to that policy, the market will rapidly come to expect it, and they won't value your country's currency this year more or less than your long-term level target says your currency will be worth. The markets will know you'll put the path back on track, and the weak form of the efficient markets hypothesis says you can't have publicly predictable price changes. That's the "rational expectations channel" which is the second key to achieving price stability. The third key is to use prediction markets on how much money you need to create, so you don't create too much or too little in the first place. But that's just icing on the cake.

Q. But what if even trying to target a nominal path still doesn't work?

A. If you (1) create increasingly more or less money as you go under or over your nominal path target, (2) show the market that you are fully committed to that target, and (3) use a prediction market to target the size of intervention, then *you will hit* 2%/year inflation, or 5%/year NGDP growth, or constant gold prices, or any other single nominal price target you choose. You could have your prices deflate by exactly 2% per year, not that you should, but you could, and the price level wouldn't blow up over or under that. That's the power of targeting a path! If you pick literally any single nominal variable you want—the price of silver, the price of a median haircut, literally anything—and you publicly declare a target price path and then consistently loosen when under the target and tighten when over the target, increasing the size of your action as you get further away from the path, and the market knows and believes this, that price path will be achieved, period. I mean, you can't say that a kilo of gold should have a constant price path of $1 because you can't afford to buy back and destroy enough dollars to make that be true. But that is the opposite of the problem that you currently have. You cannot run out of ability to make prices go higher. Running out of reserves can prevent a central bank from enforcing that its currency have a minimum value, but you cannot run out of ones and zeroes when it comes to enforcing your currency's maximum value. Just sell more of the currency!

Q. Okay, I could see firmly committing on that policy up to creating $15, but any more than that, and I'd have to give up.

A. Then the price will not be stable. The market will know that if there's enough deflation, you'll print $15 and then give up. They can see that coming, so they'll get nervous as soon as the price starts to drop. If the markets think you're not truly committed, they're very likely to test you. This is true even if you're doing something that ought to be impossible to fail at, like keeping a price high or putting a ceiling on your currency's value, because markets know that central banks often get nervous and change their minds.

You've got to be ready to actually follow through! If you are truly committed you can stabilize any one price path, but if you stop stabilizing the path, it stops being stable, and markets can make billions of dollars if they can successfully predict when you'll give up.

But remember, you literally can't lose at keeping a nominal price high so long as your determination holds. Just hold down the trigger until the target is destroyed or you run out of bullets, bearing in mind that you cannot run out of bullets.

Q. So I'd have to really commit myself? That's scary! Well, I guess I could commit myself and then change my mind later if it doesn't work out. But what if I undershoot my price path, print even more money next year, and still can't hit my targeted price path?

A. Then *you did not create enough money, god damn it*. That's what the prediction market is for. But even then, if you're consistently an embarrassing 1% under the path, that's a whole lot more stable than undershooting your 2% inflation target by a variable amount every year.

Q. What if I'm undershooting my price path by more and more each year?

A. Increase the size of your action even faster as you get even further away from the target path. Like, have your action increase as the square of the amount you missed, or something like that. Better yet, target the market forecast. Better yet, use a prediction market to figure out how much money you need to create or destroy.

Q. What if no matter how much money I print it still doesn't change prices?

A. Then why is your government even bothering to collect taxes?

Q. I mean… couldn't the problem be that the money I create is staying in bank accounts instead of doing anything?

A. Uh, aren't you currently paying positive interest on bank reserves held at your central account? If you think that's the problem you definitely should not be paying positive interest on reserves.

Q. But paying 0.25% interest on reserves makes me feel better.

A. Why?

Q. I don't know, it just does. Uh, maybe it causes my banks not to make stupid loans if they know they can at least get 0.25% interest by holding the money inside me?

A. I guess that's possible, and stupid loans are a problem, but… if you do pay positive interest on reserves, that makes banks more reluctant to lend and reflects a tighter monetary policy. It changes the amount of money that's 'enough'. It means you'll need to create even more, more money, when you've already been undershooting your inflation target for several years running. Indeed, some relatively wise central banks are charging negative interest on excess reserves.

Q. Why do you say they're only *relatively* wise?

A. Because if they were absolutely wise they'd have created enough money to create enough inflation that they wouldn't need to charge negative interest. But regardless, paying positive interest on reserves is kinda like transforming that currency into a new kind of government bond. It might be harmless if you could print enough enough-money, but otherwise it's probably not a wise thing to do, even if it makes you feel better. Think of it as a luxury reserved for good central banks that don't undershoot their inflation target.

Q. Yeah… I've been wondering if maybe my currency is pretty much interchangeable with government bonds, now. Maybe when I buy government bonds from the sort of people who own government bonds in the first place, they just keep my currency

around and don't buy anything else with it. Maybe I need to inject the money somewhere else.

A. First, like many other problems having to do with deflation, this hypothetical problem, if it existed, would be one that you could solve with *moar money*. There is some amount of money creation that overflows into the hands of people who can buy real things, that gets things moving again and causes all the factories to work at capacity and people to be employed and flows through all the trades that people can make.

Second, it does seem dubious that the injection site makes much of a difference, because demand for money is fungible, and if you add more money in one place it's generic supply that competes in a global pool of generic demand.

Third, if you think that's the problem, then for god's sake stop paying positive interest on reserves when you're already undershooting your target.

Fourth, if a wrong injection site or fungibility with bonds really was the problem, and you didn't want to brute-force it with *moar money*, you could create money and buy a broad basket of low-past-volatility equities that are easy to short-sell and hence correctly priced. Or… well, optimal monetary policy is a lot simpler than optimal politics, and I understand the latter a lot less well than the former. But maybe you could make your central bankers put on suits, and go to your politicians with hat in hand, and humbly say, "I'm sorry, we screwed up and undershot 8 years' worth of publicly declared inflation targets, and we need to give the markets a sign that we're serious this time. Can you please let us send every citizen a check for $2000 just this one time?"

Q. My government would never let me do that!

A. Are you sure? I'm not a government myself, but I kinda imagine that if you made your central bankers put on their best suits and tell the government's politicians that you just had to send all their voters a bunch of free money—

Q. But that's exactly the opposite of the way that central banks are supposed to be wiser and more paternal than governments, and tell governments 'no' about fun things they're not allowed to do! I'd never be able to look other central banks in the eye again!

A. Maybe you could compensate for the emotional damage by eating a lot of chocolate or something? It doesn't seem like the same order of problem as unemployed people slowly sinking into despair.

And if you go to the legislature and ask them for the statutory authority to mail citizens checks after undershooting your inflation mandate for three years running, the markets will go "Oh holy poot, they mean it this time, let's stop hoarding this currency" and you won't even have to mail the check.

Failing that? Bleeping do it. Think of it as declaring a country-wide dividend where every citizen gets an equal share of the new money needed to animate an economy that expands over time. I mean, I wouldn't recommend mailing too many checks because that really would make it harder to destroy money later and prevent overinflation. But it would tell the markets you really, honestly meant your new monetary stance of Bleep This Bleeping Underinflation, I Own A Printing Press.

Q. Listen, you don't understand how things are in my country! My government is full of politicians just looking for an excuse to tear me to bits, and if I so much as asked to mail checks to people, they would! Frankly, I'm worried that I couldn't print as much money as you think I should, even if I wanted to.

A. Then I guess you're not really in independent control of your country's monetary policy, huh?

Q. That's… harsh. Though, the thought of not being blamed for anything does sound nice.

A. Well, to be clear, if you aren't already printing all the money you can print, under whatever political constraints, and promising the markets to print all the money you can print later until you reach a declared path target, then you are to blame if you undershoot your inflation target. Until you're doing that, you're not doing everything you can; and whatever level of money is flowing through your economy, you did choose that particular level and no higher, so you're responsible for whatever damage it does.

Q. But… you just don't realize the obstacles here! What if other countries don't like me weakening my currency relative to their currency? They'll yell at me that I'm exporting my deflation to them!

A. If another country thinks their currency is too expensive, they can print more of it.

Q. But that just exports the deflation back to me. Isn't this a zero-sum game?

A. No. If you both print more money, then your relative exchange rate stays the same and you both undershoot your stated inflation targets less pathetically every bleeping year after year.

Q. But what if they refuse to print more money, and then blame me for adding to their deflationary pressures?

A. Then you can send them a recorded video message of you screaming at the top of your lungs that they should just *print more money*. No country that has not physically run out of ones and zeroes has the right to blame anyone or anything else for their own currency's excess value.

Q. Hm… I just realized, if they did make their currency cheaper relative to mine, they'd export more goods to me, which could steal away jobs from my country! Should I really be reminding them of that?

A. That is not how Ricardo's Law of Comparative Advantage works. Imagine a world where nobody invests in other countries or builds up foreign currency reserves from year to year. In this world, every foreign car has to be purchased with foreign currency that was obtained by selling them a domestic computer in the same year. Right? So your relative exchange rate doesn't affect the number of computers you need to sell to buy a car—the price of foreign cars in domestic computers. Conversely, if a country is selling you cars and then just keeping some of your currency in warehouses while it slowly depreciates, then they'll send you more cars than you ship computers to them, regardless of relative exchange rates.

Q. Then you're not arguing that I should print lots of money to make my currency cheaper so I can sell more things to foreigners and make sure that I'm stealing their

jobs instead of the other way around?

A. No! One, lots of the countries trying to sell you things have median incomes much lower than yours, and I frankly don't see how it could be anything but mustache-twirling cartoon villainy if I did advise you how to thrive at their expense. Two, that's not how Ricardo's Law of Comparative Advantage works. And three, if being able to get more cars by making computers and trading them for cars, actually 'destroyed jobs', then combine harvesters would also reduce employment because they produced tons more grain and destroyed farming jobs. If that was really the way economics worked, then the 100-fold increase in agricultural productivity since the medieval era when 98% of the population was made up of farmers, would have caused only 3% of your population to have jobs today. The people inside you are better off when they have more stuff, not when they need to do more work. Have you heard of the broken window fallacy?

Q. That's where, if your economy isn't doing well, you just break a lot of windows, and then people have to repair their windows, which gives employment to glaziers, who spend their wages at bakeries, who pay farmers, and the whole economy is stimulated and does better, right?

A. Well, that's the fallacy. The problem is that you're not accounting for opportunity costs. Breaking a window just means that you have to divert glass, labor, and money from other uses—the repair, and the money for the repair, aren't magical events that occur without trading off against anything else. Furthermore, the idea that the town does better, just because more money is being spent, implies that there wasn't yet enough money to animate all the trades that could be made. But if there's not enough money to go around, so that money rather than glassmaking capacity is the limiting factor on how many windows get made, then breaking windows means that limiting-factor money gets diverted from somewhere else!

I mean… the only way breaking windows could actually add jobs, is if the economy was being limited by low money flow so there was spare glassmaking capacity and spare labor, and if the person paying to repair the windows took the money out of an otherwise inactive bank account, and then the windows-repairer didn't try to save more money later to make up for the loss. Not only would that happen very rarely, but if it did happen, it would mean you'd been asleep on your job! The only real benefit is coming from spending down an otherwise inactive bank account to add money to a money-limited system! The whole purpose of your existence is to make sure that the amount of glass flowing is bound by the amount of sand and heat available, not by there being insufficient money to flow the other way.

The broken windows fallacy is a fallacy, there isn't some counterintuitive way to make people be better off by breaking their windows—or by outlawing clever harvesters that produce more grain with less labor—or by outlawing building cars by sending other people computers in exchange for cars—so long as you're creating enough money. But you shouldn't be worrying about which exact particular places your economy is allocating its labor, or whether it's more efficient to get cars by building them from scratch versus trading computers for them. I mean, maybe there should be some part of your government that worries about things like runaway occupational licensing or real marginal tax rates owing to benefit phase-outs, but not the central bank part!

Q. But my government does keep telling me to think about both inflation and jobs. That means I can't put just inflation on a particular target path like you say, because

then I wouldn't be thinking about jobs too! I need to think about two things at once, which is why I have to use an ad-hoc policy and hold big important meetings where I change my mind all the time. This dual target also explains why I'm always very worried about overshooting 2% inflation even when employment has been dropping, and why I keep ending up with too little inflation and too little employment simultaneously.

A. Yes, it's bad when people are unemployed. It's very unpleasant and it destroys real value. You want to minimize that as much as you can. This gets us into another topic, which is that even targeting inflation at a 2%/year level path isn't all that great an idea. It's better than trying to target the amount of base money, sure. But you could still have something bad happen where people became more reluctant to trade at the same time as a lot of money was destroyed. In that case you might end up with the same amount of total 'inflation' per trade, even though there were idle factories and people that you could put to work by creating more money.

Q. So I should just make more money whenever I eyeball that I think my economy is doing less than it could, but stop if inflation goes too high?

A. Honestly, that might be better than some of the other things you could be doing! Though it would be important to only stop when the market said inflation was about to go too high, not when you were feeling a little nervous that it might go high later.

However, there's a simpler and more formal answer that accomplishes the same thing and creates much more stability. Instead of targeting 2%/year inflation, target the total amount of money changing hands in your economy and make that quantity go up by 5%/year on a level price path. This quantity is called NGDP, or Nominal Gross Domestic Product.

Q. Huh, NGDP. I remember hearing bad things about that measure. Like, it's a stupid measure of my country's health, because it doesn't take into account how a computer at the same price can be much more powerful. And people even try to include military spending into NGDP, and so on.

A. That's okay! We're not using NGDP as a proxy measure of how much fun your country is having. We're using NGDP as exactly what it is, a measure of the total flow of money changing hands.

When NGDP accelerates above trend, your country's monetary policy is too loose, and there'll be too much money competing for each transaction. If NGDP drops below trend, your country's monetary policy is too tight, and there won't be enough money for each transaction.

And again, price-setters are reluctant to lower their prices, often much more so than people are reluctant to pay increased prices—or price-setters are first in line to raise prices when they have market power, but try to be last in line to drop them. It's not a rational-agent theorem, but it's definitely an empirically observed fact. So when NGDP drops, many transactions will stop happening at all, which destroys the real value of the gains from trade, which is bad. By targeting NGDP, you're targeting the flow of money directly. It means that every shopkeeper knows that the country as a whole will spend 5% more money next year, no matter what else happens. Now doesn't that sound nice and stable?

Q. What does that have to do with unemployment, though?

A. In most ways, NGDP is a mirror image of Nominal Gross Domestic Income, the amount of money that everyone receives to spend. Some economists suggest you should be targeting NGDI instead, because it seems to be a more stable estimate that gets adjusted afterwards less often than NGDP estimates. But leaving that aside and just inverting the way we look at things, keeping NGDI on a level upward path ensures that money-flow is available to pay everyone who wants to work. That's the secret hidden inside the idea that when unemployment rises, you should create more money. What you're really doing is adding more wage-flow so that more people can be employed. Think of NGDI as being like a game of musical chairs—all the people who are employed need flowing money to pay them. So if there isn't enough NGDI flowing through the system, somebody has to become unemployed! That's why sharp drops in employment track the graph for sharp drops in NGDP much more than they track the graph for lower inflation. So to make sure the economy can steadily add as many jobs as it has room for, you target the path of NGDI or the total amount of money flowing, not 'inflation'.

Q. But if I don't target inflation, won't inflation go totally out of control?

A. It would be very hard for that to happen under an NGDP level targeting regime. If the total amount of money flowing always increases at exactly 5% per year and returns to path from any level deviation, it'd be very hard for prices in the economy to regularly go up at 10% per year. I mean, maybe you're worried that a specified NGDP target implies that there'd be twice as much money-flow per transaction if half your country's residents died to a bioengineered superplague—

Q. Yes! That would be very bad. Prices could double!

A. Then you can just target NGDP per capita—not per employee, of course, but per inhabitant of the currency area—and that would be fine too. That way you won't get large amounts of inflation no matter what happens.

Q. Will that also prevent asset bubbles from forming? Sometimes people blame me for that too.

A. Any asset bubble that can happen with 5%/year NGDP growth is probably one that would've happened no matter what you did. Plus, remember, a central bank only has one price of money to control, and it affects literally everything in your country. You should use that one lever to make sure that the country as a whole will always spend 5% more money next year. So long as you do that, asset bubbles shouldn't do much damage when they pop, and you probably can't even prevent them in the first place, and if you did try to mess with them and went off the NGDP level path to do so you'd be screwing up everything else. The way your country's security regulators can prevent price bubbles is by making sure that an asset class is very cheap and easy to short—that there's no obstacles that add expense or difficulty or uncertainty to buying put options.

Not to mention, a lot of times when people yell 'bubble' the market goes up further before it eventually goes down. Which is just an ordinary problem of not knowing whether an asset is too high or too low, and you might not be any smarter about that than hedge-fund managers. But if an asset price goes too high and people can actually tell, it's often because there's a systemic difficulty that makes shorting the asset too difficult or too expensive.

Regardless, all of that just shouldn't be your concern. Just focus on making sure that 5% more nominal money will be spent next year. If a price bubble can't change that

by inflating or bursting, it probably can't affect the rest of the economy very much. And trying to 'pop' the price of one asset class you think might maybe be a 'bubble', is definitely not worth the collateral damage of allowing the whole country's NGDP to drop below trend, or the collateral damage of abandoning your declared target path.

Q. Okay, I'm not sure I believe you about this 'NGDP targeting' business, but you've at least convinced me to try, you know, printing an amount of money that strikes me as insane. And you'd better believe I am going to buy back all those bonds and destroy this money the second this economy starts to hyperinflate… huh, that's interesting.

A. What happened?

Q. Prices went up 1.5% and unemployment got closer to its normal level, though labor force participation is still down... that means it's now time to raise interest rates to above zero, right?

A. *NO!* Wait, what? What are you doing? You'll choke off your recovery! Do you have any idea how badly the markets have already reacted several years earlier because they already knew you would tighten monetary policy the instant there was any hint of recovery?

Q. But if I don't raise interest rates now, I won't have any room to cut interest rates later, if things get worse again—

A. *Aaarg! Stop!* That is not how interest rate targeting works!

Q. It's not? I always thought that cutting the interest rate stimulates the economy. And it works, so far as I can tell. Back when things were better, I'd cut rates half a point and the economy would go vssshhoooom just like pressing the accelerator pedal on a car. But I can't cut interest rates below zero, or, I mean, some people say I can, but I'm not sure I believe them. So obviously, if I might need to cut rates later in case of a recession, I should raise the interest rate now when it won't do much damage. That way I have room to cut it later!

A. *NO.* Don't. That's like stopping your antibiotic treatment in mid-course so you can 'increase the amount of antibiotics later' if you get sicker. It's not just based on an instinctive brain-level misunderstanding of whether it's 'taking antibiotics' or 'adding more antibiotics' that is the effective intervention, it actually makes the disease more resistant. Charging 3% interest means a very different thing in terms of monetary policy, depending on whether you're in Japan or Zimbabwe. In Zimbabwe, if you offer someone a loan at 3% interest, it is a super-cheap loan. In Japan, if you offer someone a loan at 3% interest, it is an expensive loan. Depending on whether there's more or less inflation, the same nominal interest rate on a loan can make the money cheap or expensive. If you raise rates now, you're tightening monetary policy, which brings the natural interest rate downward, which in the future will make the same nominal interest rate target represent a tighter monetary policy.

Q. I'm not sure I understand.

A. If you raise rates now, inflation will be lower than it otherwise would. Which means that if you later 'drop' future rates to 0.1% in case of a recession, that future 0.1% interest rate will be less effective and represent less monetary stimulus, than a rate of 0.1% would have represented if you'd just held the rate constant today. You are not giving yourself more ammunition before you hit the 'zero bound'. You are giving

yourself less ammunition before you hit the 'zero bound', and also you're choking off your economy's nascent recovery.

Q. So you're saying… raising rates now will make the economy weaker, so when I press the accelerator pedal later, it will be less effective?

A. No, more like braking changes the real distance of the accelerator from the control system's zero, which isn't the same as the accelerator's physical distance from the floor mat.

Q. What?

A. I'm saying that you need to think in terms of the effect of an interest rate instead of the effect of an interest rate drop. If you raise interest rates now, a future nominal rate of 0.1% will be less powerful in an absolute sense later. So even though you'll have the ability to 'drop' interest rates later, you'll be dropping it to a level of stimulus that's less powerful in an absolute sense, because inflation will be lower after you signal a tightened monetary policy, so the real interest rate represented by a fixed nominal rate will be higher and money will be more expensive—

Q. I still don't understand.

A. Look, just… you don't need to 'create more ammunition for later' because you can't run out of ammunition. You can just create more ones and zeroes if the economy needs them, no matter what the 'interest rate' looks like.

Q. But if I don't raise interest rates now, I'll be embarrassed in front of the other central banks!

A. If you tighten rates now then you will have to keep rates even lower for longer and create even more money later because inflation will go down and, worse, the markets will have learned that you undershoot your inflation target. They will expect overly-timid monetary policy in the future, which by the EMH (weak form) encourages traders to hoard more of your currency now and is a more contractionary monetary policy now.

If you'd cut rates fast enough in 2008, you could have kept the money flow from cratering, and a 2% nominal interest rate would represent a lower real interest rate than we have today. And then you wouldn't have needed to do quantitative easing later on. Not to mention all the banks that wouldn't have failed.

Well, it isn't any different today. The more money you print now, the less you will be embarrassed later. The more you tighten rates now, the longer you'll have to keep them 'embarrassingly low'.

Q. But interest rates have already been low for so long! It doesn't look good when I have to keep interest rates low for super-long! It signals that I think my economy is weak!

A. I'm sorry, but nobody believes your predictions about the economy any more. They've seen you overstate your predictions every year for a decade, and they have a market forecast to look at instead. Traders do care about how the central bank thinks the economy is doing, but that's because every time you think the economy is doing well, you tighten monetary policy and they want to forecast the resulting damage.

Q. But… if I raise interest rates by half a percentage point, won't it signal that I think the economy is doing really well, and people will believe me and perk up and the economy will do better and stocks will rise?

A. Is that what actually happened the last time you tried that?

Q. Well, no, but—

A. Look. You're making your life more complicated than it needs to be. If you're undershooting your target path for inflation or for whatever, create more money. Do not destroy money. Do not start undoing your quantitative easing. Do not raise interest rates. Not for signaling reasons, not to give yourself "ammunition for later", not because "interest rates have been low for too long". Just don't… tighten… policy when you're still undershooting your target. Any nominal variable you're targeting, whether it's inflation or NGDP, if you undershoot the path, then loosen monetary policy. It's *not that complicated*.

Q. Hey, I just thought of a clever idea. Maybe I can get my government to borrow more money and spend it. That way I can create less new money and still meet my inflation target!

A. Please don't. That will accomplish *literally nothing* except to create a huge burden of government debt. If every dollar the government spends is one less dollar you create, it has *literally zero* net stimulatory effect on the economy. If you think the economy as a whole is spending too little money, you should *create more money*. There is no reason the government's fiscal policy should be involved. At all. Ever. There is no way your government can push on money-flow by borrowing a dollar to spend, that you, the central bank, cannot do at lower human cost by creating a temporary dollar from scratch. The only reason for a government to buy a highway is if the highway is worth the money. Doing it for 'stimulus' is completely pointless if the central bank is effectively targeting inflation or any other nominal variable. Every dollar the government borrows to spend, is one less dollar the central bank creates to reach their nominal target. If the 'fiscal multiplier' is not zero, it means the central bank is not effectively targeting the price of anything and your president ought to be fired. Your inflation target is your problem and you shouldn't try to shove it onto your government.

Q. That's certainly an interesting political stance, but…

A. Political stance? It is an empirical statement about how money works. Remember when there was this big, scary 'sequester' that was supposed to sharply cut US government spending in the middle of a recession? The Federal Reserve scraped up some more courage, printed a corresponding amount of money, and nothing happened. There was a nice little to-do where the Keynesians predicted economic disaster and the market monetarists said 'Nothing will happen so long as the Fed prints a corresponding amount of money' and the market monetarists were experimentally correct. It was a clear-cut test of two competing theories and the results were also clear. If you are able to print more money when more money is needed, your government can cut spending without any monetary effects.

I mean, maybe you won't get a new highway and people's cars will crash. But the part where the economy crashes due to decreased total spending is something you can prevent by adding more money. In fact this will happen automatically if the central bank is in a regime where they will respond incrementally to incremental movement away from their price target—

Q. Wait… hold on… god damn it, not again. I'm sorry, I need to put this conversation on hold while I bail out one of my country's big banks.

A. What? Why would you do that? Wouldn't that create a huge moral hazard?

Q. I know that! God, I know. But if I don't bail out this bank, some other banks might fail too!

A. So?

Q. And some brokerages might fail! There'd be systemic contagion!

A. Okay…?

Q. And almost everyone has their hands in someone else's pants! Our whole financial sector could implode!

A. Truly, the tree of prudent investment must be watered from time to time with the blood of idiots. Shall we watch the fireworks together? I'll grab some popcorn.

Q. No, see… it's bad if lots of financial companies go bankrupt! I've got to stop that from happening! It's a disaster with banks in it, and I'm the Central Bank, so it must be my job to stop it!

A. Why do you even care? So a bunch of financial companies go kerplooey and vaporize the virtual money of some rich people who trusted other people wearing fancy business suits. How does that affect an average household with two children? If anything, some suckers will have less currency with which to buy yachts, and some of the steel and concrete could go into making toys for the children instead. On a more abstract level, your rich-sucker institutions would have less currency with which to make mediocre investments, and the economic capacity thus freed up could go into good investments instead.

Q. Such a huge financial disaster would have catastrophic effects on the real economy! The company that employs the parents at a haircut shop wouldn't be able to get a loan to make payroll! The rich people who were buying haircuts from them will lose a bunch of money on the stock market and be unable to afford haircuts!

A. That would be a problem of not enough money flowing, an instance of the problem class where the economy still has a bunch of factories but people don't have enough money to buy things. You can prevent this problem by creating more money.

Look, you don't even need to think about the details! Just ignore all the theatrics and keep outputting whatever amount of money the prediction market says is necessary to keep the total money flow, or Nominal Gross Domestic Product, on a level path target of 5% per year. I mean, if your country's whole financial sector is melting down, the prediction market is probably going to tell you to create a lot of money, but that's okay.

Um, it is admittedly true that this is really, really not the time to flinch from printing any required amount of money. You only want the bad banks to go bankrupt, not for all the banks to go bankrupt because there isn't enough money in the entire system to pay off even the non-stupid loans—that latter part really is your job to prevent.

Q. Do you have any idea the size of disaster we'd be risking?! What if the new money doesn't reach people fast enough!?

A. If market forecasts start to indicate that the real economy might be about to tank, I guess you could ask the legislature for the statutory authority to mail your citizens checks. If it's really going bad that fast, you can use direct deposit, even. I think the politicians would let you do it under those conditions… right? I guess you'd have to be pretty sure they'd let you do it.

Q. Even if the government did let me do it, are you seriously telling me that as a central bank, I should watch my country's entire financial sector just burn to the ground?

A. Not the entire sector. A bunch of people who made bad investments lose their shirts. If you were in a contagion environment where lots of banks were lending to bad investors or creating complicated derivatives with other banks who were, then lots of people lose their shirts. The good banks who stayed virtuous pick up the slack using all that new money you're creating—temporarily, until velocity picks up—and a few years later, everyone's investing more wisely.

As a central bank you have one job, to regularize the total flow of money through the economy. You make sure there's enough money in the total system for people to pay off good loans. You make sure that the family with two kids has enough money in hand to pay for groceries.

As a central bank, you have one superpower—the absolute ability to control the path of any one nominal variable. Use that power to promise every shopkeeper that the country will spend 5% more money next year. Promise every employee that the country will have 5% more nominal income to flow into wages, promise every honest bank that there will be 5% more money flow to pay back loans to productive enterprises. And then tell your too-big-to-fails to go hang!

You might need to sell a ludicrously huge amount of currency to keep money flow on-path if over fifty percent of your financial sector is burning merrily to the ground, especially if this is the first time you're doing this and people are scared your determination will break. And then, as good banks start to take over the loan-making capacity, you'll need to buy back a lot of that ludicrously huge amount of currency. But that's all you need to do.

Q. This sounds really scary! I'd rather take a bunch of ad-hoc half-measures and throw enormous amounts of money at a bunch of total bastards and create terrifying amounts of moral hazard just to avoid the temporary cleansing fire that would have cleaned a lot of gunk out of the system and maybe even reduced inequality!

A. Well, you certainly won't be alone if you go that route.

Q. How can you just say that I should let a huge amount of virtual money go up in smoke while temporarily creating enough new money to prevent real-world disruption? How can I let so many banks fail when everyone with more than $250,000 invested at a bad bank will only have $250,000 left plus whatever percentage of the bank's other assets were recoverable, with a compensating amount of new money being temporarily added somewhere else to level the total money flow through the economy?

A. It's simple! You just declare an NGDP path target, set up a prediction market, tell an intern to do whatever the prediction market says, and then go on vacation for the rest of your life as a central bank!

Q. This is madness!

A. Madness? *This... is... market monetarism!*

Originally posted to social media on February 11, 2016.

Outline of an approach to AGI Estimation

We are worried about what will happens when we make a system that can do the important things that human can do, like programming and science. Will it explode off into infinity as it finds better ways to improve itself or will it be a slower more manageable process? There are a number of productive ways we can react to this, not limited to:

1. Attempt to make AI systems controllable by focusing on the control problem
2. Attempt to predict when we might get agi by looking at progress, e.g. ai impacts
3. Attempt to predict what happens when we get artificial general intelligence by looking at current artificial intelligences and current general intelligences and making predictions about the intersection.
4. Figure out how to make intelligence augmentation, so we can improve humans abilities to compete with full AGI agents.

This blog post is looking at the third option. It is important, because the only reason we think AGIs might exist is the existence proof of humans. Computer AGI might be vastly different in capabilities but we can't get any direct information about them, our only possible approach to predicting their capability is to adjust our estimate of what humans can do based on the difference between humans and narrow AI on different tasks and those tasks importance for economically important generality.

There are three core questions to AGI Estimation that I have identified so far.

1. What do we mean by 'general intelligence'
2. How important is 'general intelligence' to different economic activities
3. How good are narrow AIs at parts of the processes involved in general intelligence vs humans. Can we estimate how good an AGI would be at another task based on the comparison?

Generality

We often speak and act as if there is something to one person being better than another person and mental acts in general. We try and measure it with IQ, we think there is something there. However there are a number of different things "general intelligence" might actually be each of which might lead to better performance, while still being general in some way.

1. Better at tasks in general
2. Better at absorbing information by experimentation, so that per bit of information the task performance improves quicker
3. Better at holding more complex behaviours, so that the limit of task performance is higher

4. Better at understanding verbal and cultural information, so that they can improve tasks performance by absorbing the information about tasks acquired by other people.

5. Better in general at figuring out the relevant information and how it should be formatted/collected, so they can select the correct paradigm for solving a problem.

Number 1 seems trivially untrue. So we shouldn't worry about a super intelligent AGI system automatically being better at persuading people, unless it has been fed the right information. You don't know what the right information is for a super intelligence to get good at persuading, so probably still worth while being paranoid.

My bet is that "generally better" is a normally a mixture of aspects 2-4 (which feed into each other), but different aspects will dominate in different aspects of economic activity. Let us take physics for an example, I suspect that a humans potential is more limited by 4, the ability to read and understand other people research, now that we have super computers that can be programmed so that complex simulations can be off loaded from the human brain (so 3 is less important). If we throw more processing power and compute at the "general intelligence" portion of physics and this mainly improve 4, then we should expect it to get up to speed on the state of the art of physics a lot more quickly than humans, but not necessarily get better with more data. If 5 can be improved in "general" then we should expect a super intelligent AGI physicist to have a completely different view point on physics and experimentation in short order to humans.

We can look at this by looking at task performance in humans, if someone can be a lot better at absorbing information over a large number of tasks, then we should expect AGIs to be able to vary a lot on this scale. If humans general task performance variation depends heavily on being able to access the internet/books, then we don't have good grounds for expecting computers to go beyond humanities knowledge easily, unless we get indication from narrow AI that it is better at something like absorbing information.

Before getting into the other question I will illustrate what I mean by the different sorts of generality. I'll build off a simple model of human knowledge that Matt Might used to explain what grad school is like can be found here. We can imagine different ways an AGI might be better than a human at filling in the circle.

They could be fast at filling in the circle without humans help (2), be able to keep more of the circle in mind at once (3). They could be fast at filling in the circle with human culture (4a) or be able to fill more of the circle from human culture (4b). If they are faster at doing PhD level work, in general, they will expand the knowledge more (5). At some point I will put in-line diagrams or animations of the differences.

Knowing which of these facets humans change upon most and have the most impact on ability to get things done, seems pretty important.

Improving Generality and Economic Activity

However we are not interested in all tasks we are mainly interested in the ones that lead to greater economic activity and impact upon the world.

So we need to ask how important human brains are in economic activity in general, to know how much impact improving them would have.

For example for programming a computer (an important task in recursive self-improvement), how big a proportion of the time spent in the task is the humans intelligence vs total task. For example when an AI researcher is trying to improve an algorithm how much time of making a new algorithm variant is needed for thinking vs running the algorithm with different parameters. For example if it takes 3 days for a new variant of AlphaGo Zero, to train itself and be tested and 1 day for the human to analyse the result and tweak the algorithm, speeding up the human portion of the process 1000 fold won't have a huge impact. This is especially the case if the AGI is not appreciably better at aspect 2 of 'generalness', it will still need as many iterations of running AlphaGo Zero as a human to be able to improve it.

The impact of what are we improving

This brings us to the question, when we create super human narrow AI systems which aspect of the system are super to human. Is it the ability to absorb knowledge, the amount of knowledge the system has been given (as it can be given knowledge more quickly than humans) or is it able to hold a more complex model than we can hold in our heads?

If we determine that the aspects that are super human are those that are very important for different economic activities and the aspects that we have not improved are not important, then we should increase our expectation of the impact of AGI. This should hopefully not make us too provincial in our predictions about what AGI should be capable of doing.

A first stab at the kind of reasoning that I want to attempt is here where I attempt to look at how quickly Alpha Go Zero learns compared to humans based on number of games played. This might be a bad comparison, but if it is we should try and improve upon it.

Conclusion

I have outlined a potential method for gaining information about AGI that relies on mainly experiments on human task performance.
We do not have lots of options when it comes to getting information about AGI so we should evaluate all possible ways of getting information.

Fighting the evil influence of Facebook (but keeping the good bits): a manifesto and how-to guide

This is a linkpost for https://thomas-sittler.github.io/facebook/

In defence of epistemic modesty

This piece defends a strong form of epistemic modesty: that, in most cases, one should pay scarcely any attention to what you find the most persuasive view on an issue, hewing instead to an idealized consensus of experts. I start by better pinning down exactly what is meant by ‘epistemic modesty’, go on to offer a variety of reasons that motivate it, and reply to some common objections. Along the way, I show common traps people being inappropriately modest fall into. I conclude that modesty is a superior epistemic strategy, and ought to be more widely used - particularly in the EA/rationalist communities.

[gdoc]

Provocation

I argue for this:

> In virtually all cases, the credence you hold for any given belief should be dominated by the balance of credences held by your epistemic peers and superiors. One’s own convictions should weigh no more heavily in the balance than that of one other epistemic peer.

Introductions and clarifications

A favourable motivating case

Suppose your mother thinks they can make some easy money day trading blue-chip stocks, and plans to kick off tomorrow shorting Google on the stock market, as they’re sure it’s headed for a crash. You might want to dissuade her in a variety of ways.

You might appeal to an outside view:

> Mum, when you make this short you’re going to be betting against some hedge fund, quant, or whatever else. They have loads of advantages: relevant background, better information, lots of data and computers, and so on. Do you really think you’re odds on to win this bet?

Or appeal to some reference class:

> Mum, I’m pretty sure the research says that people trying to day-trade stocks tend not to make much money at all. Although you might hear some big successes on the internet, you don’t hear about everyone else who went bust. So why should you think you are likely to be one of these remarkable successes?

Or just cite disagreement:

> Look Mum: Dad, sister, the grandparents and I all think this is a really bad idea. Please don’t do it!

Instead of directly challenging the object level claim (i.e. “Google isn’t overvalued, because X”). These considerations attempt to situate the cogniser within some population, and from characteristics of this population infer the likelihood of this cogniser getting things right.

Call the practice of using these techniques considerations epistemic modesty. We can distinguish two components:

1. ‘In theory’ modesty: That considerations of this type should in principle influence our credences.
2. ‘In practice’ modesty: That one should in fact use these considerations when forming credences.

Weaker and stronger forms of modesty

Some degree of modesty is (almost) inarguable. If one leaves for work on Tuesday and finds all your neighbours left their bins out, that’s at least reason to doubt your belief bins were on Thursday, and perhaps sufficient to believe instead bins are on Tuesday (and follow suit with your bins). If it appears that, say, the coagulation cascade ‘couldn’t evolve’, the near unanimity of assent for evolution among biologists at least counts against this, if not a decisive reason, despite one's impressions, that it could. Nick Beckstead suggests something like ‘elite common sense’ forms a prior which one should be hesitant to diverge from without good reason.

I argue for something much stronger (c.f. the Provocation above): in theory, one’s credence in some proposition P should be almost wholly informed by modest considerations. That, ceteris paribus, the fact it appears to you that P should weigh no more heavily in one’s determination regarding P than knowing that it appears to someone else that P. Not only is this the case in theory, but it is also the case in practice. One’s all things considered judgement on P should be just that implied by an idealized expert consensus on P, no matter one’s own convictions regarding P.

Motivations for more modesty

Why believe ‘strong form’ epistemic modesty? I first show families of cases where ‘strong modesty’ leads to predictably better performance, and show these results generalise widely.[1]

The symmetry case

Suppose Adam and Beatrice are perfect epistemic peers, equal in all respects which could bear on them forming more or less accurate beliefs. They disagree on a particular proposition P (say “This tree is an Oak tree”). They argue about this at length, such that all considerations Adam takes to favour “This is an Oak tree” are known to Beatrice, and vice versa.[2] After this, they still disagree: Adam has a credence of 0.8, Beatrice 0.4.

Suppose an outside party (call him Oliver) is asked for his credence of P, given Adam and Beatrice’s credences and their epistemic peer-hood to one another, but bereft of any object-level knowledge. He should split the difference between Adam and Beatrice

- 0.6: Oliver doesn't have any reason to favour Adam over Beatrice's credence for P as they are epistemic peers, and so splitting the difference gives the least expected error.[3] If he was faced with a large class of similar situations (maybe Adam and Beatrice get into the same argument for Tree 2 to Tree 10,000) Oliver would find that difference splitting has lower error than biasing to either Adam or Beatrice's credence.

Adam and Beatrice should do likewise. They also know they are epistemic peers, and so they should also know that for whatever considerations explain their difference (perhaps Adam is really persuaded by the leaf shapes, but Beatrice isn't) Adam's take and Beatrice's take are no more likely to be right than one another. So Adam should go (and Beatrice vice-versa), "I don't understand why Beatrice isn't persuaded by the leaf shapes, but she expresses the same about why I find it so convincing. Given she is my epistemic peer, 'She's not getting it', and, 'I'm not getting it' are equally likely. So we should meet in the middle".

The underlying intuition is one of symmetry. Adam and Beatrice have the same information. The correct credence regarding P given this information should not depend on which brain Adam or Beatrice happens to inhabit. Given this, they should hold the same credence[4], and as they Adam is as likely to be further from the truth than Beatrice, the shared credence should be in the middle.

Compressed sensing of (and not double-counting) the object level

It seems odd that both Adam and Beatrice do better discarding their object level considerations regarding P. If we adjust the scenario above so they cannot discuss with one another but are merely informed of each other's credences (and that they are peers regarding P), the right strategy remains to meet in the middle.[5] Yet how come Adam and Beatrice are doing better if they ignore relevant information? Both Adam and Beatrice have their 'inside view' evidence (i.e. what they take to bear on the credence of P) and the 'outside view' evidence (what each other think about P). Why not use a hybrid strategy which uses both?

Yet to whatever extent Adam or Beatrice's hybrid approach leads them to diverge from equal weight, they will do worse. Oliver can use the 'meet in the middle strategy' to get an expectedly better accuracy than either biasing towards their own inside view determination. In betting terms, Oliver can arbitrage any difference in credence between Adam and Beatrice.

We can explain why: the credences Adam and Beatrice offer can be thought of as very compressed summaries of the considerations they take to bear upon P. Whatever 'inside view' considerations Adam took to bear upon P are already 'priced in' to the credence he reports (ditto Beatrice). Modesty is not ignoring this evidence, but weighing it appropriately: if Adam then tries to adjust the outside view determination by his own take on the balance of evidence, he double counts his inside view: once in itself, and once more by including his credence as weighing equally to Beatrice's in giving the outside view.

One's take on the set of considerations regarding P may err, either by bias,[6] ignorance, or 'innocent' mistake. Splitting the difference between you and your peer's very high level summary of these captures the great fraction of benefit of hashing out where these summaries differ.[7] Modesty correctly diagnoses that one's high level

summary is no more likely to be more accurate than one's peers, and so holds those in equal regard, even in cases where the components of one's own summary are known better.

Repeated measures, brains as credence censors, and the wisdom of crowds

Modesty outperforms non-modesty in the n=2 case. The degree of outperformance grows (albeit concavely) as n increases.

Scientific fields often have to deal with unreliable measurement. They commonly mitigate this by having repeat measurement. If you have a crummy thermometer, repeating readings several times improves accuracy over just the once. Human brains also try and measure things, and they are also often unreliable. It is commonly observed that nonetheless the average of their measurement tends to lie closer to the mark than the vast majority of individual measurements. Consider the commonplace 'guess how many skittles are in this jar' or similar estimation games: the usual observation is that the average of all the guesses is better than all (or almost all) the individual guesses.

A toy model makes this unsurprising. The individual guesses will form some distribution centered on the true value. Thus the expected error of a given individual guess is the standard deviation of this distribution. The expected error of the average of all guesses is given by the standard error, which is the standard deviation divided by root(number of guesses):[8] with 10 individuals, the error is about 3 times smaller than the expected error of each individual guess; with 100, 10 times smaller; and so on.

Analogously, human brains also try to measure credences or degrees of belief, and are similarly imperfect to when they're trying to estimate 'number of X'. Yet one may expect a similar effect to this 'wisdom of crowds' to operate here too. In the same way Adam and Beatrice would do better in the situation above if they took the average (even if it went against their view of the balance of reasons by their lights), if Adam-to-Zabaleta (all epistemic peers) investigated the same P, they'd expect to do better if they took the average of their group versus steadfastly holding to the credence they arrived at 'by their lights'. Whatever inaccuracies that may throw off their individual estimates of P somewhat cancel out.

Deferring to better brains

The arguments above apply to cases where one is an epistemic peer. If not, one needs to adjust by some measure of 'epistemic virtue'. In cases where Adam is an epistemic superior to Beatrice, they should meet closer to Adam's view, commensurate with the degree of epistemic superiority (and vice versa).

Although reasons for being an epistemic superior could be 'they're a superforecaster' or 'they're smarter than me', perhaps the most common source of epistemic superiors lie under the heading of 'subject matter expert'. On topics from human nutrition, to voting rules, to the impact of the minimum wage, to the nature of consciousness, to basically anything that isn't trivial, one can usually find a fairly large group of very smart people who spend many years studying that topic, who make public their views

about this topic (sometimes not even behind a paywall). That they at least have a much greater body of relevant information and have spent longer thinking about it gives them a large advantage compared to you.

In such cases, the analogy might be that your brain is a sundial, whilst theirs is an atomic clock. So if you have the option of taking their readings rather than yours, you should do so. The evidence a reading of a sundial provides about the time conditional on the atomic clock reading is effectively zero. ‘Splitting the difference’ in analagous epistemic cases should result with both you and your epistemic superior agreeing that they are right and you are wrong.

Inference to the ideal epistemic observer

We can summarise these motivations by analogy to ideal observers (used elsewhere in perception and ethical theory). We can gesture that an ideal (epistemic) observer is just that which is able to form the most accurate credence for P given whatever prior: we can explain they have vast intelligence, full knowledge of all matters that bear upon P, perfect judgement, and in essence all epistemic virtues in excelsis.

Now consider this helpful fiction:

> The epistemic fall: Imagine a population solely comprised of ideal observers, who all share the same (correct) view on P. Overnight their epistemic virtues are assailed: they lose some of their reasoning capacity; they pick up particular biases that could throw them one way or another; they lose information, and so on, and each one to varying degrees.
>
> They wake up to find they now have all sorts of different credences about P, and none of them can remember what credence they all held yesterday. What should they do?

It seems our fallen ideal observers can begin to piece together what their original credence was about P by finding out more about their credences and remaining epistemic virtue, and so backpropagate their return to epistemic apotheosis. If they find they’re all similarly virtuous and are evenly scattered, their best guess is the ideal observer was in the middle of the distribution (c.f. the wisdom of crowds). If they see a trend that those with greater residual virtue tend to hold a higher credence in P, they should attempt to extrapolate this trend to suggest the ideal agent origin from which they were differentially blown of course from. If they see one group demonstrates a bias that others do not, they can correct the position of this group before trying these procedures. If they find the more virtuous agents are more scattered regarding P, (or that they segregate into widely dispersed aggregations), this should make them very unsure about where the ideal observer initially was. And so on.

Such a model clarifies the benefit of modesty. Although we didn’t have some grand epistemic fall, it is clear we all fall manifestly short of an ideal observer. Yet we all fall short in different respects, and in different degrees. One should want to believe whatever one would believe if one was an ideal observer, shorn of one’s manifest epistemic vices. Purely immodest views must say their best guess is the ideal observer would think the same as they do, and hope that all the vicissitudes of their epistemic vice happen to cancel out. By accounting for the distribution of cognisers, modesty allows a much better forecast, and so a much more accurate belief. And the

best such forecast is the strong form of modesty, where one’s particular datapoint, in and of itself, should not be counted higher than any other.

Excursus: Against common justifications for immodesty

So much for strong modesty in theory. How does it perform in practice?

One rough heuristic for strong modesty is this: for any question, find the plausible expert class to answer that question (e.g. if P is whether to raise the minimum wage, talk to economists). If this class converges on a particular answer, believe that answer too. If they do not agree, have little confidence in any answer. Do this no matter whether one’s impression of the object level considerations that recommend (by your lights) a particular answer.

Such a model captures all the common sense cases of modesty - trust the results in typical textbooks, defer to consensus in cases like when to put the bins out, and so on. I now show it is also better in many cases where people think it is better to be immodest.

Being ‘well informed’ (or even true expertise) is not enough

A common refrain is that one is entitled to ‘join issue’ with the experts due to one having made some non-trivial effort at improving one's knowledge of the subject. “Sure, I accept experts widely disagree on macro-economics, but I’m confident in neo-Keynesianism after many months of careful study and reflection.”

This doesn’t fly by the symmetry argument above. Our outsider observes widespread disagreement in the area of macroeconomics, and that many experts who spend years on the subject nonetheless greatly disagree. Although it is possible the ideal observer would have been in one or another of the ‘camps’ (the clustering implies intermediate positions are less plausible), the outsider cannot adjudicate which one if we grant the economists in each appear to have similar levels of epistemic virtue. The balance of this outside view changes imperceptibly if another person who despite a few months of study remains nowhere near peerhood (let alone superiority) of these divided experts, happens to side with one camp or another. By symmetry, one's own view of the balance of reason should remain unchanged if this ‘another person’ happened to be you.

The same applies even if you are a bona fide expert. Unless the distribution of expertise is such that there is a lone ‘world authority’ above all others (and you’re them) your fellow experts form your epistemic peer group. Taking the outside view is still the better bet: the consensus of experts tends to be right more often than dissenting experts, and so some difference splitting (weighed more to the consensus owing to their greater numbers) is the right answer.[9]

Common knowledge ‘silver bullet arguments’

Suppose one takes an introductory class in economics. From this, one sees there must be a ‘knock-down’ argument against a minimum wage:

> Well, suppose you’re an employee whose true value on the free market is less than the minimum wage. But under the minimum wage, the firm might not decide on charitably employing above your market value, and just firing you instead. You’re worse off, as you’re on the dole, and the firm’s worse off, as it has to meet its labour demand another way. Everyone’s lost! So much for the minimum wage!

Yet one quickly discovers economists seem to be deeply divided over the merits of the minimum wage (as they are about most other things). See for example this poll suggesting 38 economic experts in the US are pretty evenly divided on whether the minimum wage would ‘hit’ employment for low-skill workers, and leant in favour of the minimum wage ‘all things considered’.

It seems risible to suppose these economists don’t know their economics 101. What seems much more likely is that they know other things that you don’t which make the minimum wage more reasonable than your jejune understanding of the subject suggests. One need not belabour which side the outside view strongly prefers.

Yet it is depressingly common for people to confidently hold that view X or Y is decisively refuted by some point or another, notwithstanding the fact this point is well known to the group of experts that nonetheless hold X or Y. Of course in some cases one really has touched on the decisive point the experts have failed to appreciate. More often, one is proclaiming that one is on the wrong side of the Dunning-Kruger effect.

Debunking the expert class (but not you)

To the litany of cases where (apparent) experts screwed up, we can add verses without end. So we might be inclined to debunk a particular ‘expert consensus’ due to some bias or irrationality we can identify. Thus, having seen there are no ‘real’ experts to help us, we must look at the object level case.

The key question is this: “How are you better?” And it is here that debunking attempts often flounder:

An undercutting defeater for one aspect of epistemic superiority for the expert class is not good enough. Maybe one can show the expert class has a poor predictive track record in their field. Unless one has a better track record in their field, this puts you on a par with respect to this desideratum of epistemic virtue. They likely have others (e.g. more relevant object-level knowledge) that should still give them an edge, albeit attenuated.

An undercutting defeater that seems to apply equally well to oneself as the expert class also isn’t enough. Suppose (say) economics is riven by ideological bias: why are you less susceptible to these biases? The same ideological biases that might plague professional economists may also plague amateur economists, but the former retain other advantages.

Even if a proposed debunking is ‘selectively toxic’ to the experts versus you, it still might be your epistemic superior all things considered. Both Big Pharma and Professional Philosophy may be misaligned, but perhaps not so much to be orthogonal

or antiparallel to the truth: in both they still expectedly benefit by finding drugs that work or making good arguments respectively. They may still fare better overall than, “Intelligent layperson who’s read extensively”, even if they are not subject to ‘publish or perish’ or similar.

Even if a proposed debunking shows one as decisively superior to that expert class, there may be another expert class which remains epistemically superior to you. Maybe you can persuasively show professional philosophers are so compromised on consciousness that they should not be deferred to about it. Then the real expert class may simply switch to something like ‘intelligent people outside the academy who think a lot about the topic’. If it’s the case that this group of people do not share your confidence in your view, it seems outsiders should still reject it - as should you.

It need not be said that the track record for these debunking defeaters is poor. Most crackpots have a persecution narrative to explain why the mainstream doesn’t recognise or understand them, and some of the most mordant criticisms of the medical establishment arise from those touting complementary medicine. Thus ‘explaining away’ expert disagreement may not put one in a more propitious reference class than one started from. One should be particularly suspicious of debunking(s) sufficiently general that the person holding the unorthodox view has no epistemic peers - they are akin to Moses, descending from Mt. Sinai, bringing down God-breathed truth for the rest of us.[10]

Private evidence and pet arguments

Suppose one thinks one is in receipt of a powerful piece of private evidence: maybe you’ve got new data or a new insight. So even though the experts are generally in the right, in this particular case they are wrong because they are unaware of this new consideration.

New knowledge will not spread instantaneously, and that someone can be ‘ahead of the curve’ comes as no surprise. Yet many people who take themselves to have private evidence are wrong: maybe experts know about it but don’t bother to discuss it because it is so weak, or it is already in the literature (but you haven’t seen it), or it isn’t actually relevant to the topic, or whatever else. Most mavericks who take themselves to have new evidence that overturns consensus are mistaken.

The natural risk is people tend to be too partial to their pet arguments or pet data, and so give them undue weight, and so one’s ‘insider’ perceptions should perhaps be attenuated by this fact. I suspect most are overconfident here.[11] If this private evidence really is powerful, one should expect it to be persuasive to members of this expert class once they become aware of it. So it seems the credence one should have is the (appropriately discounted) forecast of what the expert class would think once you provide them this evidence.

The natural test of the power of this private evidence is to make it public. If one observes experts (or just epistemic peers) shift to your view, you were right about how powerful this evidence was. If instead one sees a much more modest change in opinion, this should lead one to downgrade your estimate as to how powerful this evidence really is (and perhaps provide calibration data for next time). Holding instead this really is decisive evidence leads one to the problematic ‘common knowledge silver bullet’ case discussed above. Inferring from this experts just can’t understand

your reasoning or are biased against outsiders or whatever else produces a suspiciously self-serving debunking argument, also discussed above.

Objections

So much for the case in favour. What about the case against? I divide objections into those ‘in theory’, and those ‘in practice’.

In theory

There’s no pure ‘outside view’[12]

It is not the case you can bootstrap an outside view from nothing. One needs to at least start with some considerations as to what makes one an epistemic peer or superior, and probably some minimal background knowledge of ‘aboutness’ to place topics under one or another expert class.

In the same way large amounts of our empirical information are now derived by instrument rather than direct application of our senses (but were ultimately germinated from direct sensory experience), large amounts of our epistemic information can be derived by deferring to better (or more) brains rather than using our own, even if this relies on some initial seed epistemology we have to realise for ourselves. This ‘germinal set of claims’ can still be modestly revised later.

Immodestly modest?

One line of attack from the social epistemology literature is that strong forms of modesty are self-defeating. If one is modest, one should assumedly be modest about ‘What is the right way to form beliefs if epistemic peers disagree with you?’ Yet one finds that very few people endorse the sort of epistemic modesty advocated above. When one looks among potential expert classes, such as more intelligent friends of mine (i.e. friends of mine), epistemologists, and so on, conciliatory views like these command only a minority. So the epistemically modest should vanish as they defer to the more steadfast consensus.

If so, so much the worse for modesty. I offer a couple of incomplete defences:

One is haggling over the topic of disagreement. In my limited reading of ‘equal weight/conciliatory views and their detractors’, I take the detractors to be suggesting something like “one is ‘within one’s rights’ to be steadfast”, rather than something like “you’re more accurate if you’re steadfast”. Maybe there are epistemic virtues which aren’t the same as being more accurate. Yet there may be less disagreement on ‘conditional on an accuracy first view, is modesty the right approach?’

This only gets so far (after all, shouldn’t we be modest whether only to care about accuracy?) A more general defence is this: the ‘what if you apply the theory to itself?’ problem looks pretty pervasive across theories.[13] Accounts of moral uncertainty that in whatever sense involve weighing normative theories by their plausibility tend to run into problems if the same accounts are applied ‘one level up’ to meta-moral uncertainty. Bayesian accounts of epistemology seem to go haywire if we think one

should have a credence in Bayesian epistemology itself, especially if one assigns any non-zero credence on any theory which entails object level credences have undefined values.

Closer to home, milder versions of conciliation (e.g. “Pay some attention to peer disagreement, but it’s not the only factor”) share a similarly troublesome recursive loop (“Well, I see most other people are steadfast, so I should update to be a bit less conciliatory, but now I have to apply my modified view to this disagreement again”) and neat convergence is not guaranteed. The theories which avoid this problem (e.g. ‘Wholly steadfast, so peer disagreement should be ignored’), tend to be the least plausible on the object level (e.g. That if you believe bins are on Thursday, the fact all your neighbours have their bins out on Tuesday is not even reason to reconsider your belief).

A solution to these types of problems remains elusive. Yet modesty finds itself in fairly good company. It may be the case that a good resolution to this type of issue would rule out the strong form of modesty advocated here, in favour of some intermediate view. Until then, I hope the (admittedly inelegant) “Be modest, save for meta-epistemic norms about modesty itself” is not too great a cost to bear across the scales from the merits of the approach.

In practice

I take most of the action to surround whether modesty makes sense as a practical procedure in the real world, even granting it’s ‘in theory’ virtue. Given the strength of modesty, I advocate, the fact we use something like it in some cases, and we can identify it can help in others, is not enough. It needs to be shown as a better strategy than even slightly weaker forms, in circumstances deliberately selected to pose the greatest challenge to strong modesty.

Trivial (and less trivial) non-use cases

For some topics there’s no relevant epistemic peers or superiors to consider. This is commonly the case with pretty trivial beliefs (e.g. my desk is yellow).

Modesty also doesn’t help much for individual tastes, idiosyncrasies, or circumstances. If Adam works best listening to Bach and Beatrice to Beethoven, they probably won’t do better ‘meeting in the middle’ and both going half-and-half for each (or maybe picking a composer intermediate in history, like Mozart). Anyway, Adam is probably Beatrice’s significant epistemic superior on “What music does Adam work best listening to?”, and vice-versa. One can also be credulous of claims like “It turned out this diet really helped my back pain”: perhaps it’s placebo, or perhaps it is one of those cases where different things work for different people, and one expects in such cases individuals to have privileged access to what worked for them.[14]

There will be cases where one really is plowing a lonely furrow where there aren’t any close epistemic peers or superiors. It’s possible I really am the world’s leading expert on “How many counter-factual DALYs does a doctor avert during their career?”, because no one else has really looked into this question. My current role involves investigating global catastrophic biological risks, which appears understudied to the point of being pre-paradigmatic.

These comprise a very small minority of topics I have credences about. Yet even here modesty can help. One can use more distant bodies of experts: I am reassured that my autumnal estimate for the ‘DALY question’ coheres with expert consensus that medical practice had a minor role in improvements to human health, for example. Even if I don’t have any epistemic peers, I can simulate some by asking, “If there were lots of people as or more reasonable than me looking at this, would I expect them to agree with my take?” Given that the econometric-esque methods I deploy to the answer the ‘DALY question’ could probably be done better by an expert, and in any case reasonable people are often sceptical of these in other areas, I am less confident of my findings than my ‘inside view’ suggests, which I take to be a welcome corrective to ‘pet argument’ biases.[15]

In theory, the world should be mad

Whether devoured by Moloch, burned by Ra, trapped by aberrant signalling equilibria, or whatever else, we can expect to predict when apparent expert classes (and apparent epistemic peers) are going to collectively go wrong. With this knowledge, we can know which topics we should expect to ourselves to outperform expertise. Rather than the scenario where we commonly find ourselves looking up (at experts) or around (at our peers), we find ourselves in many situations where those who are usually epistemic peers or superiors are below us - and above us, only sky.

We could distinguish two sorts of madness, a surprising absence of expertise and a surprising error of expertise:

The former is a gap in the epistemic market. Although an important topic should be combed over by a body of experts, for whatever reason it isn’t, and so it takes surprisingly little effort to climb to the summit of epistemic superiority. In such cases our summaries of expert classes as ranging over a broad area conceal the degree of expertise is very patchy: public health experts generally know a great deal about the health impacts of smoking; they usually know much less about the health impacts of nicotine.

The latter is a stronger debunking argument. One appeals to some features of the world that generates expertise and suggests that these expertise generating features are anti-correlated to the truth, thus one can adjudicate between warring expert camps (or just indict all so-called ‘experts’) based on this knowledge. One strong predictor of incompatibilism regarding free will among is believing in God. If we are confident these beliefs in God are irrational, then we can winnow the expert class by this consideration and side with the compatibilist camp much more strongly.

Yet, similar to the problems of debunking mentioned earlier, that there is a good story suggesting one of these things does not imply one will do better ‘striking out on one's own’. Even in cases of disease where accuracy is poorly correlated to expert activity, it is hard to think of cases where these line up orthogonal or worse. Big pharma studies are infamous, but even if you’re in big pharma optimising for ‘can I get evidence to support my product’, your drug actually working does make this easier. Even in pre-replication crisis psychology, true results would be overrepresented versus false ones in the literature compared to some base rate across generated hypotheses.

The ‘residual’ expert class still often remains better. Although most public health experts know little about nicotine per se, there are some nearby health experts, perhaps scattered across our common-sense demarcation of fields, who do know about the impacts of nicotine. It may still take quite a lot of effort to reach parity or

superiority to these. Even if we want to strike all theists from free will philosophers, compatibilism does not rise close to unanimity, and so cautions against extremely high confidence this is the correct view.[16] So, I aver, the world is not that mad.

Empirically, the world is mad

One can offer a more direct demonstration of world madness, and so modesty: outperformance.

A common reply is to point to a particular case where those being modest would have gotten it wrong. There are lots of cases where amateurs and mavericks were ridiculed by common sense or experts-at-the-time, only to be subsequently vindicated.

Another problem is the modest view introduces a lag - it seems one often needs to wait for the new information to take root among one's epistemic peers before changing one's view, whilst a cogniser just relying on the object level updates on correct arguments 'at first sight'. It is often crucially important to be fast as well as right in both empirical and moral matters: it is extremely costly if a view makes one slower to recognise (among many other past moral catastrophes) the horror of slavery.

Yet modesty need not infallible, merely an improvement. Citing cases where it goes poorly is (hopefully less than) half the story. Modesty does worse in cases the maverick is right, yet better where the maverick is wrong: there are more cases of the latter than the former. Modesty does worse in being sluggish in responding to moral revolutions, yet better at avoiding being swept away by waves of mistaken sentiment: again, the latter seem more common than the former.[17]

Maybe one can follow a strategy such that you can 'pick the hits' of when to carve out exceptions, and so have a superior track record. Yet, empirically, I don't see it. When I look at people who are touted as particularly good at being 'correct contrarians', I see at best something like an 'epistemic venture capitalist' - their bold contrarian guesses are right more often than chance, but not right more often than not. They appear by my lights to be unable to judiciously 'pick their battles', staking out radical views in topics where there isn't a good story as to why the experts would be getting this wrong (still less why they're more likely to get it right). So although they do get big wins, the modal outcome of their contrarian take is a bust.[18]

Modesty should price in the views of better-than-chance contrarians into how it weighs consensus. Confidence in a consensus view should fall if a good contrarian takes aim at it, but not so much one now takes the contrarian view oneself. If one happens to be a particularly successful contrarian one should follow the same approach: "I get these right surprisingly often, but I'm still wrong more often than not, so it might be worth it to look into this further to see if I can strike gold, but until then I should bank on the consensus view."

Expert groups are seldom in reflective equilibrium

Even if modesty works well in the ideal case of a clearly identified 'expert class', it can get a lot messier in reality:

1. Suppose one is in the early 1940s and asks, “Is there going to be explosives with many orders of magnitude more power than current explosives?” One can imagine if one consulted explosive experts (however we cash that out), their consensus would generally say ‘no’. If one was able to talk to the physicists working on the Manhattan project, they would say ‘yes’. Which one should an outside view believe?[19]
2. Most people believe god exists (the so called ‘common consent argument for God’s existence’); if one looks at potential expert classes (e.g. philosophers, people who are more intelligent), most of them are Atheists. Yet if one looks at philosophers of religion (who spend a lot of time on arguments for or against God’s existence), most of them are Theists - but maybe there’s a gradient within them too. Which group, exactly, should be weighed most heavily?

So constructing the ideal ‘weighted consensus’ modesty recommends deferring to can become a pretty involved procedure. One must carefully divine whether a given topic lies closer to the magisterium of one or another putative expert class (e.g. maybe one should lean more to the physicists, as the question is really more ‘about physics’ than ‘about explosives’). One might have to carefully weigh up the relevant epistemic virtues of various expert classes that appear far from reflective equilibrium from one another (so perhaps one might use likely selection effect of philosophy of religion party discount the apparent support this provides). One might have to delve into complicated issues of independence: although most people may believe god exists, unlike guesses of how many skittles are in the jar, they are not all forming this belief independently from one another.[20]

This exercise begins to look increasingly insider-view-esque. Trying to determine the right magisterium involves getting closer to object level considerations about ‘aboutness’ of topics; trying to tease apart issues of independence and selection amount to looking at belief forming practices, and veer close to object level justifications for the belief in question. At some point it becomes extraordinarily challenging to try and back-trace from all these factors to the likely position of the ideal observer: the degrees of freedom these considerations invite (and the challenge in estimating them reliably) make strong modesty go worse.

One should not give up too early, though: modesty can still work pretty well even in these tricky cases. One can ask whether there’s any communication between the classes, and if so any direction of travel (e.g. did some explosive experts end up talking to the physicists, and agreeing they were right? Vice-versa?), even if they were completely isolated, one can ask if a third group having access to both made a decision (e.g. the agreement of the U.S. and German governments with the implied view of the physicists). This is a lot more involved, but the expected ‘accuracy yield per unit time spent’ may still be greater than (for example) making a careful study of the relevant physics.

A broader modification would be ‘immodest only for the web of belief, but modest for the weights’: one uses an inside view to piece together the graph of considerations around P, but one still defers to consensus on the weights. This may avoid cases where (for example) strong modesty may mistake astronomers as the expert class for about space travel being infeasible (versus primordial rocket scientists), even though astronomers and rocket scientists agreed about the necessary acceleration, but astronomers were inexpert on the key question as to whether that explanation could be produced.[21]

What if one cannot even do that? Then modestly (rightly) offers a counsel of despair. If an area is so fractious there's no agreement, with no way to see which of numerous of disparate camps have better access the truth of the matter; so suffused with bias that even those with apparent epistemic virtues (e.g. judgement, intelligence, subject-matter knowledge) cannot be seen to even tend towards the truth; what hope does one have to do better than they? In attempting to thread the needle through these hazards towards the right judgement, one will almost certainly run aground somewhere or somehow, alike all one's epistemic peers or superiors who made the attempt before. Perhaps reality obliges us to undertake these doxastic suicide missions from time to time. If modesty cannot help us, it can at least provide the solace of a pre-emptive funeral, rather than (as immodest views would) cheer us on to our almost certain demise.

Somewhat satisfying Shulman

Carl Shulman encourages me to offer my credences and rationale in cases he takes to be particularly difficult for my view, and suggests in these cases I either arrive at absurd credences or I am covertly abandoning the strong modesty approach. I offer these below for readers to decide - with the rider that if these are in fact absurd, 'I'm an idiot' is a competing explanation to 'strong modesty is a bad epistemic practice' (and that, assuredly, whatever one's credence on the latter, one's credence in the former should be far greater).

Proposition (roughly); **Credence (ish)**; (Modesty-based) rationale, in sketch

Theism; **0.1**[22]; Mostly discount common consent (non-independence) and PoR (selection). Major hits from more intelligent people/ better informed tend to be atheist, but struggle to extrapolate this closer to 0 given existence proofs of very epistemically virtuous religious people.

Libertarian free will; **0.1**; Commands a non-trivial minority across virtuous epistemic classes (philosophers, intelligent people, etc), only somewhat degraded by selection worries.

Jesus rose from the dead; **0.005**; Christianity in particular a very small fraction of possibility space of Theism. Support from its widespread support is mostly (but not wholly) screened off by non-independence effects. Relevant (but distant) expert classes in history etc. weigh adversely.

There has been a case of cold fusion; **10^-5**; Strong pan scientific consensus against, cold fusion community looks renegade and much less epistemically virtuous. Base rate of these conditional on no effect gives very adverse reference class.

ESP; **10^-6**; Very strong (but non-complete) trophism among elite common sense, scientists, etc; bad predictive track records for ESP researchers; distant consensuses highly adverse. Some greatly attenuated boost from survey data/small fraction of reasonable believers.

Practical challenges to immodesty

Modesty can lead to double-counting, or even groupthink. Suppose in the original example Beatrice does what I suggest and revise their credences to be 0.6, but Adam doesn't. Now Charlie forms his own view (say 0.4 as well) and does the same

procedure as Beatrice, so Charlie now holds a credence of 0.6 as well. The average should be lower: (0.8+0.4+0.4)/3, not (0.8+0.6+0.4)/3, but the results are distorted by using one-and-a-half helpings of Adam's credence. With larger cases one can imagine people wrongly deferring to hold consensus around a view they should think is implausible, and in general the nigh-intractable challenge from trying to infer cases of double counting from the patterns of 'all things considered' evidence.

One can rectify this by distinguishing 'credence by my lights' versus 'credence all things considered'. So one can say "Well, by my lights the credence of P is 0.8, but my actual credence is 0.6, once I account for the views of my epistemic peers etc." Ironically, one's personal 'inside view' of the evidence is usually the most helpful credence to publicly report (as it helps others modestly aggregate), whilst ones all things considered modest view usually for private consumption.

Community benefits to immodesty

Modesty could be parasitic on a community level. If one is modest, one need never trouble oneself with any 'object level' considerations at all, and simply cultivate the appropriate weighting of consensuses to defer to. If everyone free-rode like that, no one would discover any new evidence, have any new ideas, and so collectively stagnate.[23] Progress only happens if people get their hands dirty on the object-level matters of the world, try to build models, and make some guesses - sometimes the experts have gotten it wrong, and one won't ever find that out by deferring to them based on the fact they usually get it right.[24]

The distinction between 'credence by my lights' versus 'credence all things considered' allows the best of both worlds. One can say 'by my lights, P's credence is X' yet at the same time 'all things considered though, I take P's credence to be Y'. One can form one's own model of P, think the experts are wrong about P, and marshall evidence and arguments for why you are right and they are wrong; yet soberly realise that the chances are you are more likely mistaken; yet also think this effort is nonetheless valuable because even if one is most likely heading down a dead-end, the corporate efforts of people like you promises a good chance of someone finding a better path.

Scott Sumner seems to do something similar:

> In macro, it's important for people like me to always search for the truth, and reach conclusions about economic models in a way that is independent of the consensus model. In that way, I play my "worker ant" role of nudging the profession towards a greater truth. But at the same time we need to recognize that there is nothing special about our view. If we are made dictator, we should implement the consensus view of optimal policy, not our own. People have trouble with this, as it implies two levels of belief about what is true. The view from inside our mind, and the view from 20,000 miles out in space, where I see there is no objective reason to favor my view over Krugman's.

Despite this example, maybe it is the case that 'having a creative brain which makes big discoveries' is anticorrelated to 'having a sober brain well-calibrated to its limitations compared to others': anecdotally, eccentric views among geniuses are common. Maybe for most it isn't psychologically tenable to spend one's life investigating a renegade view one thinks ultimately is likely a dead-end, and in fact people do groundbreaking research generally have to be overconfident to do the best

science. If so, we should act communally to moderate this cost, but not celebrate it as a feature.

Not everyone has to do be working on discovering new information. One could imagine a symbiosis between eccentric overconfident geniuses whose epistemic comparative advantage is to who gambol around idea-space to find new considerations, and well-calibrated thoughtful people whose comparative advantage is in soberly weighing considerations to arrive at a well callibrated all-things-considered view.

Conclusion: a pean, and a plea

I have argued above for a strong approach to modesty, one which implies - at least in terms of 'all things considered view' - one's view of the object level merits counts for very little. Even if I am mistaken about the ideal strength of modesty, I am highly confident both the EA and rationalist communities err in the 'insufficiently modest' direction. I close on these remarks.

Rationalist/EA exceptionalism

Both communities endure a steady ostinato of complaints about arrogance. They've got a point. I despair of seeing some wannabe-iconoclast spout off about how obviously the solution to some famously recondite issue is X and the supposed experts who disagree obviously just need to better understand the 'tenets of EA' or the sequences. I become lachrymose when further discussion demonstrates said iconoclast has a shaky grasp of the basics, that they are recapitulating points already better-discussed in the literature, and so forth.[25]

To stress (and to pre-empt), the problem is not, "You aren't kowtowing appropriately to social status!" The problem is considerable over-confidence married with inadequate understanding. This both looks bad to outsiders,[26] but it also is bad as the individual (and the community itself) could get to the truth faster if they were more modest about their likely position in the distribution of knowledge about X, and then did commonsensical things to increase it.

Consider Gell-Mann amnesia (via Michael Crichton):

> You open the newspaper to an article on some subject you know well. In Murray's case, physics. In mine, show business. You read the article and see the journalist has absolutely no understanding of either the facts or the issues. Often, the article is so wrong it actually presents the story backward—reversing cause and effect. I call these the "wet streets cause rain" stories. Paper's full of them.
>
> In any case, you read with exasperation or amusement the multiple errors in a story, and then turn the page to national or international affairs, and read as if the rest of the newspaper was somehow more accurate about Palestine than the baloney you just read. You turn the page, and forget what you know.

Gell-Mann cases invite inferring adverse judgements based on extrapolating from in instance of poor performance. When experts in multiple different subjects say the same thing (i.e. Murray and Crichton chatted to an expert on Palestine who had the same impression), this adverse inference gets all the stronger.

I think we have information some to many pieces of work or corporate projects in our community share this property: that although it might look good or groundbreaking to us as relatively less-informed, domain experts in the fields it touches upon tend to report the work is misguided or rudimentary. Although it is possible to indict all these judgements, akin to a person who gives very adverse accounts of all of their previous romantic partners, we may start to wonder about a common factor explanation. Our collective ego is writing checks our epistemic performance (or, in candour, performance generally) cannot cash; general ignorance, rather than particular knowledge, may explain our self-regard.

To discover, not summarise

It is thought that to make the world go better new things need to be discovered, above and beyond making sound judgements on existing knowledge. Quickly making accurate determinations of the balance of reason for a given issue is greatly valuable for the latter, but not so much for the former.

Yet the two should not be confused. If one writes a short overview of a subject ‘for internal consumption’ which gives a fairly good impression of what a particular view should be, one should not be too worried if a specialist complains that you haven’t covered all the topics as adequately as one might. However, if one is aiming to write something which articulates an insight or understanding not just novel to the community, but novel to the world, one should be extremely concerned if domain experts review this work and say things along the lines of, “Well, this is sort of a potted recapitulation of work in our field, and this insight is widely discussed”.

Yet I see this happen a lot to things we tout as ‘breakthrough discoveries’. We want to avoid case where we waste our time in unwitting recapitulation, or fail to catch elementary mistakes. Yet too often we license ourselves to pronounce these discoveries without sufficient modesty in cases where there’s already a large expert community working on similar matters. This does not preclude these discoveries, but it cautions us to carefully check first. On occasions where I take myself to have a new insight in areas outside my field (most often philosophy), I am extremely suspect of my supposed discovery: all too often would this arise from my misunderstanding, or already be in the literature somewhere I haven’t looked. I carefully consult the literature as best as I can, and run the idea by true domain experts, to rule out these possibilities.[27]

Others seem to lack this modesty, and so predictably err. More generally, a more modest view of ‘intra-community versus outside competence’ may also avoid cases of having to reinvent the wheel (e.g. that scoring rule you spent six months deriving for a karma system is in this canonical paper), or for an effort to derail (e.g. oh drat, our evaluation provides worthless data because of reasons we could have known from googling ‘study design’).

Paradoxically pathological modesty

If the EA and rationalist communities comprised a bunch of highly overconfident and eccentric people buzzing around bumping their pet theories together, I may worry about overall judgement and how much novel work gets done, but I would at grant this at least looks like fertile ground for new ideas to be developed.

Alas, not so much. What occurs instead is agreement approaching fawning obeisance to a small set of people the community anoints as ‘thought leaders’, and so centralizing on one particular eccentric and overconfident view.[28] So although we may preach immodesty on behalf of the wider community, our practice within it is much more deferential.

I hope a better understanding of modesty can get us out of this ‘worst of both worlds’ scenario. One, it can at least provide better ‘gurus’ to defer to. Modesty also helps in correcting the overly wide gap we have between our gurus and other experts, and the overly narrow gap between ‘intelligent layperson in the community’ and ‘someone able to contribute to the state of the art on a topic of interest. Some topics are really hard: being able to become someone with ‘something useful to say’ about these not take days but take years; there are many deep problems we must concern ourselves with; that the few we select as champions, despite their virtue, cannot do them all alone; and that we need all the outside help we can get.

Coda

What the EA community mainly has now is a briar-patch of dilettantes: each ranges widely, but with shallow roots, forming whorls around others where it deems it can find support. What it needs is a forest of experts: each spreading not so widely; forming a deeper foundation and gathering more resources from the common ground; standing apart yet taller, and in concert producing a verdant canopy.[29] I hope this transformation occurs, and aver modesty may help effect it.

Acknowledgements

I thank Joseph Carlsmith, Owen Cotton-Barratt, Eric Drexler, Ben Garfinkel, Roxanne Heston, Will MacAskill, Ben Pace, Stefan Schubert, Carl Shulman, and Pablo Stafforini for their helpful discussion, remarks, and criticism. Their kind help does not imply their agreement. The errors remain my own.

---

[1] Much of this follows discussion in the social epistemology literature about conciliationism, or the ‘equal weight view’. See here for a summary

[2] They also argue at length about the appropriate weight each of these considerations should have on the scales of judgement. I suggest (although this is not necessary for this argument) that in many cases most of the action lies in judging the ‘power’ of evidence. In most cases I observe people agree that a given consideration C influences the credence one holds in P; they usually also agree in its qualitative direction; the challenge comes in trying to weigh each consideration against the others, to see which considerations one’s credence over P should pay the greatest attention to.

This may represent a general feature of webs of belief being dense and many-many (A given credence is influenced by many other considerations, and forms a consideration for many credences in turn), or it may simply be a particular feature of webs of belief in which humans perform poorly: although I am confident I can determine the sign of a particular consideration, I generally don’t back myself to hold credences (or likelihood ratios) to much greater precision than the first significant digit, and I (and, perhaps,

others) struggle in cases where large numbers of considerations point in both directions.

[3] In the literature this is called ‘straight averaging’. For a variety of technical reasons this doesn’t quite work as a peer update rule. That said, given things like bayesian aggregation remain somewhat open problems, I hope readers will accept my promissory note that there will be a more precise account which will produce effectively the same results (maybe ‘approximately splitting the difference’) through the same motivation.

[4] C.f. Aumann’s agreement theorem. As an aside (which I owe to Carl Shulman), straight averaging will not work in some degenerate cases where (similar to ‘common knowledge puzzles’) one can infer precise observations from the probabilities stated. The neatest example I can find comes from Hal Finney (see also):

> Suppose two coins are flipped out of sight, and you and another person are trying to estimate the probability that both are heads. You are told what the first coin is, and the other person is told what the second coin is. You both report your observations to each other.
>
> Let's suppose that they did in fact fall both heads. You are told that the first coin is heads, and you report the probability of both heads as 1/2. The other person is told that the second coin is heads, and he also reports the probability as 1/2. However, you can now both conclude that the probability is 1, because if either of you had been told that the coin was tails, he would have reported a probability of zero. So in this case, both of you update your information away from the estimate provided by the other.

[5] To motivate: Adam and Beatrice no longer know whether or not reasons they hold for or against P are private evidence or not. Yet (given epistemic peerhood), they have no principled reason to suppose “I know something that they don’t” is more plausible than the opposite. So again they should be symmetrical.

[6] (On which more later) it is worth making clear that the possibility of bias for either Adam or Beatrice doesn’t change the winning strategy on expectation. Say Adam’s credence for P is in fact biased upwards by 0.4. If Adam knows this, he can adjust and become unbiased, if Oliver or Beatrice knows this (and knows Adam doesn’t), the break the peerhood for Adam but can simulate unbiased Adam* which would remain a peer, and act accordingly. If none of them know this, then it is the case that Beatrice wins, as does Oliver following a non-averaging ‘go with Beatrice’ strategy. Yet this is simply epistemic luck: without information, all reasonable prior distribution candidates of (Adam’s bias - Beatrice’s bias) are symmetrical about 0.

[7] Another benefit of modesty is speed: Although it is the case Adam and Beatrice’s credence (and thus the average) gets more accurate if they have time to discuss it, and so catch one another if they make a mistake or reveal previously-private evidence, averaging is faster and the trade-off in time for better precision may not be worth it. It still remains the case, as per the first example, that they still do better, after this discussion, if they meet in the middle on residual disagreement.

[8] A further (albeit minor and technical) dividend is that although individual guesses may form any distribution (for which the standard deviation may not be a helpful summary), the central limit theorem applies to the average of guesses distribution, so it tends to normality.

[9] Even if one is the world authority, there should be some deference to lesser experts. In cases where the world expert is an outlier, one needs to weigh up numbers versus (relative) epistemic superiority to find the appropriate middle.

[10]

> God from the Mount of Sinai, whose gray top
>
> Shall tremble, he descending, will himself
>
> In Thunder Lightning and loud Trumpets sound
>
> Ordaine them Lawes…

- Milton, Paradise Lost

[11] I take the general pattern that strong modesty usually immures one from common biases is a further point in its favour.

[12] I owe this to Eric Drexler

[13] A related philosophical defence would point out that the self-undermining objection would only apply to whether one should believe modesty, not whether modesty is in fact true.

[14] I naturally get much more sceptical if that person then generalises from this N=1 uncontrolled unblinded crossover trial to others, or takes it as lending significant support against some particular expert consensus or expertise more broadly: “Doctors don’t know anything about back pain! They did all this rubbish but I found out all anyone needs to do is cut carbs!”

[15] It also provokes fear and trembling in my pre-paradigmatic day job, given I don’t want the area to have strong founder effects which poorly track the truth.

[16] For example:

> One of the easiest hard questions, as millennia-old philosophical dilemmas go. Though this impossible question is fully and completely dissolved on Less Wrong, aspiring reductionists should try to solve it on their own.

[17] Aside: A related consideration is ‘optimal damping’ of credences, which is closely related to resilience. Very volatile credences may represent the buffeting of a degree of belief by evidence large relative to one's prior - but it may also represent poor calibration in overweighing new evidence (and vice versa). The ‘ideal’ response in terms of accuracy is given by standard theory. Yet it is also worth noting that’s one prudential reasons may want to introduce further lag or lead, akin to the ‘D’ or ‘I’ components of a PID controller. In large irreversible decisions (e.g. career choice) it may be better to wait a while after one’s credences support a change to change action; for case of new moral consideration it may be better to act ‘in advance’ for precautionary principle-esque reasons.

[18] (Owed to Will MacAskill) There’s also a selection effect: of a sample of ‘accurate contrarians’, many of these may be lucky rather than good.

[19] I owe this particular example to Eric Drexler, but similar counter-examples along these lines to Carl Shulman.

[20] Another general worry is these difficult-to-divine considerations offer plenty of fudge factors - both to make modesty get the ‘right answer’ in historical cases, and to fudge present areas of uncertainty to get results that accord with one’s prior judgement.

[21] I owe both this modification and example to discussions with Eric Drexler. There are some costs - one may think there are cases one should defer to an outside view on the web of belief (E.g. Christian apologist: “Sure, I agree with scientific consensus that it’s improbable Jesus rose naturally from the dead, but the key argument is whether Jesus rose supernaturally from the dead. So the consensus for philosophers of religion is the right expert class.”) The balance of merit overall is hard to say, but such a modification still looks like pretty strong modesty.

[22] In conversation I recall a suggestion by Shulman such a credence should change one’s behaviour regarding EA - maybe one should do theology research in the hope of finding a way to extract infinite value etc. Yet the expert class for action|Theism gives a highly adverse prior: virtually no actual theists (regardless of theological expertise, within or outside EA) advocate this.

[23] I understand a similar point is raised in economics regarding the EMH and the success of index funds. Someone has to do the price discovery.

[24] I owe this mainly to Ben Pace, Andrew Critch argues similarly.

[25] For obvious reasons I’m reluctant to cite specific examples. I can offer some key words for the sort of topics I see this problem as endemic: Many-worlds, population ethics, free will, p-zombies, macroeconomics, meta-ethics.

[26] C.f. Augustine, On the Literal Meaning of Genesis:

> Usually, even a non-Christian knows something about the earth, the heavens, and the other elements of this world, about the motion and orbit of the stars and even their size and relative positions, about the predictable eclipses of the sun and moon, the cycles of the years and the seasons, about the kinds of animals, shrubs, stones, and so forth, and this knowledge he hold to as being certain from reason and experience. Now, it is a disgraceful and dangerous thing for an infidel to hear a Christian, presumably giving the meaning of Holy Scripture, talking nonsense on these topics; and we should take all means to prevent such an embarrassing situation, in which people show up vast ignorance in a Christian and laugh it to scorn.

[27] I’m uncommonly fortunate that for me such domain experts are both nearby and generous with their attention. Yet this obstacle is not insurmountable. An idea (which I owe to Pablo Stafforini) is that a contrarian and a sceptic of the contrarian view could bet on whether a given expert, on exposure to the contrarian view, would change their mind as the contrarian predicts. S may bet with C: “We’ll pay some expert $X to read your work explicating your view, if they change their mind significantly in favour (however we cash this out) I’ll pay the $X, if not, you pay the $X.

[28] C.f. Askell’s and Page’s remarks on ‘buzz’.

[29] Perhaps unsurprisingly, I would use a more modest ecological metaphor in my own case. In reclaiming extremely inhospitable environments, the initial pioneer organisms die rapidly. Yet their corpses sustain detritivores, and little by little, an initial ecosystem emerges to be succeeded by others. In a similar way, I hope that the

detritus I provide will, after a fashion (and a while), become the compost in which an oak tree grows.

Things you should never do

Things you should never do

A great article on a common fallacy among programmer-types

Contra double crux

Summary: CFAR proposes double crux as a method to resolve disagreement: instead of arguing over some belief B, one should look for a crux (C) which underlies it, such that if either party changed their mind over C, they would change their mind about B.

I don't think double crux is that helpful, principally because 'double cruxes' are rare in topics where reasonable people differ (and they can be asymmetric, be about a considerations strength rather than direction, and so on). I suggest this may diagnose the difficulty others have noted in getting double crux to 'work'. Good philosophers seem to do much better than double cruxing using different approaches.

I aver the strengths of double crux are primarily *other* epistemic virtues, pre-requisite for double crux, which are conflated with double cruxing itself (e.g. it is good to have a collaborative rather than combative mindset when disagreeing). Conditional on having this pre-requisite set of epistemic virtues, double cruxing does not add further benefit, and is probably inferior to other means of discussion exemplified by good philosophers. I recommend we look elsewhere.

What is a crux?

From Sabien's exposition, a crux for some belief B is another belief C which if one changed one's mind about C, one would change one's mind about B. The original example was the impact of school uniforms concealing unhelpful class distinctions being a crux for whether one supports or opposes school uniforms.

A double crux is a particular case where two people disagree over B and have the same crux, albeit going in opposite directions. Say if Xenia believes B (because she believes C) and Yevgeny disbelieves B (because he does not believe C), then if Xenia stopped believing C, she would stop believing B (and thus agree with Yevgeny) and vice-versa.

How common are cruxes (and double cruxes)?

I suggest the main problem facing the 'double crux technique' is that disagreements like Xenia's and Yevgeny's, which can be eventually traced to a single underlying consideration, are the exception rather than the rule. Across most reasonable people on most recondite topics, 'cruxes' are rare, and 'double cruxes' (roughly) exponentially rarer.

For many recondite topics I think about, my credence it in arises from the balance of a variety of considerations pointing in either direction. Thus whether or not I believe 'MIRI is doing good work', 'God exists', or 'The top marginal tax rate in the UK should be higher than its current value' does not rely on a single consideration or argument, but rather its support is distributed over a plethora of issues. Although in some cases undercutting what I take as the most important consideration would push my degree of belief over or under 0.5, in other cases it would not.

Thus if I meet someone else who disagrees with me on (say) whether God exists, it would be remarkable if our disagreement hinges on (for example) the evidential

argument of evil, such that if I could persuade them of its soundness they would renounce their faith, and vice versa. Were I persuaded the evidential argument from evil 'didn't work', I expect I would remain fairly sceptical of god's existence; were I to persuade them it 'does work', I would not be surprised if they maintained other evidence nonetheless makes god's existence likely on the total balance of evidence. And so on and so forth for other issues where reasonable people disagree. I suspect a common example would be reasonably close agreement on common information, yet beliefs diverging based on 'priors', comprised of a melange of experiences, gestalts, intuitions, and other pieces of more 'private' evidence.

Auxiliary challenges to double crux

I believe there are other difficulties with double crux, somewhat related to the above:

Crux-asymmetry

As implied above, even in cases where there is a crux C for person X believing B, C may not be a crux for B for person Y, but it might be something else (A?). (Or, more generally, X and Y's set of cruxes are disjoint). A worked example:

Carl Shulman and I disagree about whether MIRI is doing good research (we have money riding on it). I expect if I lose the bet, I'd change my mind substantially about the quality of MIRI's work (i.e. my view would be favourable rather than unfavourable). I don't see this should be symmetrical between Shulman and I. If he lost the bet, he may still have a generally favourable view of MIRI, and a 'crux' for him maybe some other evidence or collection of evidence.

X or Y may simply differ in the resilience of their credence in B, such that one or the other's belief shifts more on being persuaded on a particular consideration. One commoner scenario (intra-EA) would be if one is trying to chase 'hits', one is probably more resilient to subsequent adverse information than the initial steers that suggested a given thing could be hit.

A related issue is when one person believes they are in receipt of a decisive consideration for or against B their interlocutor is unaware of.

'Changing one's mind' around p=0.5 isn't (that) important

In most practical cases, a difference between 1 and 0.51 or 0 and 0.49 is much more important than between 0.49 and 0.51. Thus disagreements over confidence of dis/belief can be more important, even if they may not count as 'changing one's mind': I probably differ more with a 'convinced Atheist' than an a 'doubter who leans slightly towards Theism'.

Many arguments and considerations are abductive, and so lend strength to a particular belief. Thus a similar challenge applies to proposed cruxes - they may regard the *strength*, rather than *direction*, of a given consideration. One could imagine the 'crux' between me and the hypothetical convinced Atheist is they think that the evidential problem of evil provides overwhelming disconfirmation for Theism, whilst I think its persuasive, perhaps decisive, but not so it drives reasonable credence in Theism down to near-zero.

Sabien's exposition recognises this, and so suggests one can 'double crux' over varying credences. So in this sample disagreement, the belief is 'Atheism is almost certain', and the crux is 'the evidential argument from evil is overwhelming'. Yet our language for credences is fuzzy, and so what would be a crux for the difference between (say) 'somewhat confident' versus 'almost certain' hard to nail down in a satisfactory inter-subjective way. An alternative where a change of raw credence is 'changing ones mind' entails all considerations we take to support our credence in a given a belief are cruxes.

Intermezzo

I suggest these difficulties may make a good diagnosis for why double cruxing has not always worked well. Anecdata seems to vary from those who have found it helpful to those who haven't seen any benefit (but perhaps leaning towards the latter), and remarks along the lines of wanting to see a public example.

Raemon's subsequent exegesis helpfully distinguishes between the actual double crux technique. and "the overall pattern of behaviour surrounding this Official Double Crux technique". They also offer a long list of considerations around the latter which may be pre-requisite for double cruxing working well (e.g. Social Skills, Actually Changing your Mind, and so on).

I wonder what value double crux really adds, if Raemon's argument is on the right track. If double cruxing requires many (or most) of the pre-requisites suggested, all disagreements conditioned on meeting these pre-requisites will go about as well whether one uses double crux or some other intuitive means of subsequent discussion.

A related concern of mine is a 'castle-and-keep' esque defence of double crux which arises from equivocating between double crux *per se* and a host of admirable epistemic norms it may rely upon. Thus when defended double crux may transmogrify from "look for some C which if you changed your mind about you'd change your mind about B too" to a large set of incontrovertibly good epistemic practices "It is better to be collaborative rather than combative in discussion, and be willing to change ones mind, (etc.)" Yet even if double cruxing is associated with (or requires) these good practices, it is not a necessary condition for them.

Good philosophers already disagree better than double cruxing

> *To find fault, easy; to do better, difficult*
>
> - Plutarch (paraphrased)

Per Plutarch's remark, any shortcomings in double crux may count for little if it is the 'best we've got'. However, I believe I can not only offer a better approach, but this approach already exists 'in the wild'. I have the fortune of knowing many extraordinary able philosophers, and not only observe their discussions but (as they also have extraordinary reserves of generosity and forbearance) participate in them as well. Their approach seems to do much better than reports of what double cruxing accomplishes.

What roughly happens is something like this:

1. X and Y realize their credences on some belief B vary considerably.
2. X and Y both offer what appear (to their lights) the strongest considerations that push them to a higher/lower credence on B.
3. X and Y attempt to prioritize these considerations by the sensitivity of credence in B is to each of these, via some mix of resilience, degree of disagreement over these considerations, and so forth.
4. They then discuss these in order of priority, moving topics when the likely yield drops below the next candidate with some underlying constraint on time.

This approach seems to avoid the 'in theory' objections I raise against double crux above. It seems to avoid some of the 'in practice' problems people observe:

- These discussions often occur (in fencing terms) at double time, and thus one tends not to flounder trying to find 'double-cruxy' issues. Atheist may engage Theist on attempting to undermine the free will defence to the argument from evil, whilst Theist may engage Atheist on the deficiencies of moral-antirealism to prepare ground for a moral argument for the existence of god. These may be crux-y but they may be highly assymetrical. Atheist may be a compatibilist but grant libertarian free will for the sake of argument, for example: thus Atheist's credence in God will change little even if persuaded the free will defence broadly 'checks out' if one grants libertarian free will, and vice versa.
- These discussions seldom get bogged down in fundamental disagreements. Although Deontologist and Utilitarian recognise their view on normative ethics is often a 'double crux' for many applied ethics questions (e.g. Euthanasia), they mutually recognise their overall view on normative ethics will likely be sufficiently resilient such that either of them 'changing their mind' based on a conversation is low. Instead they turn their focus to other matters which are less resilient, and thus they anticipate a greater likelihood of someone or other changing their mind.
- There appears to be more realistic expectations about the result. If Utilitarian and Deontologist do discuss the merits of utilitarianism versus (say) kantianism, there's little expectation of 'resolving their disagreement' or that they will find or mutually crucial considerations. Rather they pick at a particular leading consideration on either side and see whether it may change their confidence (this is broadly reflected in the philosophical literature: papers tend to concern particular arguments or considerations, rather then offering all things considered determinations of broad recondite philosophical positions).
- There appear to be better stopping rules. On the numerous occasions where I'm not aware of a particularly important consideration, it often seems a better use of everyone's time for me to read about this in the relevant literature rather than continuing to discuss (I'd guess 'reading a book' beats 'discussing with someone you disagree with' on getting more accurate beliefs about a topic per unit time surprisingly often).

Coda: Wherefore double crux?

It is perhaps possible for double crux to be expanded or altered to capture the form of discussion I point to above, and perhaps one can recast all the beneficial characteristics I suggest in double crux verbiage. Yet such a program appears a fool's errand: the core of the idea of double crux introduced at the top of the post is distinct from generally laudable epistemic norms (c.f. Intermezzo, supra), but also the practices of the elite cognisers I point towards in the section above. A concept of double crux so altered to incorporate these things is epiphenomenal - the engine which is driving the better disagreement is simply those other principles and practices double crux has now appropriated, and its chief result is to add terminological overhead, and, perhaps, inapt approximation.

I generally think the rationalist community already labours under too much bloated jargon: words and phrases which are hard for outsiders to understand, and yet do not encode particularly hard or deep concepts. I'd advise against further additions to the lexicon. 'Look for key considerations' captures the key motivation for double crux better than 'double crux' itself, and its meaning is clear.

The practices of exceptional philosophers set a high bar: these are people selected for, and who practice heavily, argument and disagreement. It is almost conceivable that they are better than this than even the rationalist community, notwithstanding the vast and irrefragable evidence of this group's excellence across so many domains. Double crux could still have pedagogical value: it might be a technique which cultivates better epistemic practices, even if those who enjoy excellent epistemic practices have a better alternative. Yet this does not seem the original intent, nor does there appear much evidence of this benefit.

In the introduction to double crux, Sabien wrote that the core concept was 'fairly settled'. In conclusion he writes:

> We think double crux is super sweet. To the extent that you see flaws in it, we want to find them and repair them, and we're currently betting that *repairing and refining double crux* is going to pay off better than try s*omething totally different*. [emphasis in original]

I respectfully disagree. I see considerable flaws in double crux, which I don't think have much prospect of adequate repair. Would that time and effort be spent better looking elsewhere.

Why no total winner?

Why doesn't a single power rule the world today?

[I'm taking advantage of the new "LW posts as blog posts" format to post something I'm pretty unsure about. I'm working from my memories of the blog posts, and from a discussion I had with Robin Hanson and Katja Grace in late 2012. Please let me know if any of this is inaccurate!]

One of the key differences of opinion in the Hanson-Yudkowsky AI-Foom Debate is about the idea of a "decisive advantage". If I'm not misrepresenting the parties horribly, the idea is that some point in a world with AGI, some AGI-enabled party uses their greater intelligence to increase their general power: money, resources, control over others, intelligence and suchlike. That greater power increases their ability to gain power, resulting in a snowball effect that ends with some party having control over the outcomes for all of Earth-originating life.

Robin asks the very reasonable question: if that's how things work, why hasn't it already happened? What stops the largest business using its decisive power over smaller ones to defeat and absorb them, growing ever larger and more powerful until all other businesses fall to its power? Why do we have multiple nations today, when this model would seem to predict that a single state should ultimately conquer and rule all? I don't remember Robin proposing an answer of his own: a mechanism or theoretical model that would lead us to expect multiple powers. But it seems like a good question, and it's bugged me ever since.

I think I'd need to be much more of a student of history than I am to have any confidence in an answer, so let me share some wild speculation that might at least start discussion:

- Regulation: Such growth isn't an option for legal businesses at all, because states exist. So powerful a business would challenge the power of the state, and the state is in a position to disallow that. The explicit purpose of monopoly legislation is to stop a business which has become very powerful in one area from leveraging that to become powerful elsewhere.

- Principal-agent problems: it sure would be easier to keep an empire together if you could reliably appoint generals and rulers who always did what you told them to. Especially if the round-trip time for getting them a message is on the order of weeks, and you have to entrust them with the discretion to wield tremendous power in the mean time.

- Moral norms: Nuclear weapons gave the USA a decisive advantage at the end of WWII. If the USA had been entirely ruthless and bent on power at any cost, it would immediately have used that advantage to cripple all rivals for world superpower and declared its rulership of the world.

I don't expect any of these factors to limit the growth of an AGI. Is there some more general limit to power begetting power that would also affect AGI?

The Typical Sex Life Fallacy

[Related to: Different Worlds.]

[Please note that this post contains explicit discussion of sexuality (without pictures), including discussion of my own sex life.]

[I have no moderation control, but I would definitely really appreciate it if further discussion of whether this post and ones like it are appropriate for LW move to here.]

My friend Andrew Rettek remarked to me a while back about the tremendous diversity in how people shower.

People may take anywhere between five minutes and forty minutes to shower. They may wash their hair daily, once a week, or not at all. They may wash their bodies thoroughly, only clean the parts that look dirty, only clean certain parts (such as the armpits or genitals), or just stand under the water. They may use a loofah, a sponge, or nothing. They may bring in a comb to comb out the conditioner. They may sing. They may zone out. They may jerk off. They may bathe instead, and bathing may involve reading a book or bath bombs or lighting candles and drinking a nice bottle of wine or bubble bath or none of those things at all. The one thing that is consistent is that everyone thinks the way they shower is the way normal people shower.

The reason for all this diversity, of course, is that after early childhood we don't shower together (except in locker rooms or as a form of sexual foreplay, both of which are likely to be unusual) and we rarely discuss exactly how we shower. We can get a certain amount of information about typical showers (such as length) from living with people, but again most people don't live with that many people, and the people they live with may be unusual. The rule follows: for things that are private and rarely discussed, there may be a good deal of unacknowledged diversity.

Sex is interesting because, while private, it is often discussed. People (including myself) have a certain tendency to deduce what sex is like for everyone from what sex is like for ourselves. As an example, consider pubic hair. There are innumerable thinkpieces about the pressure experienced by women to shave their pubic hair and the disgust of their male sexual partners if they are unshaven.

This has never been my experience. I have literally never had a man offer any opinion on my pubic hair whatsoever. If he did I would consider him to be an utter boor, I would never hook up with him again, and I would complain to my friends and expect all my friends to be sympathetic. My local norm is that, while of course one may have preferences about pubic hair grooming or genital size or coloring or some other traits, it is incredibly rude to voice any opinion about others' genitals other than "happy to be here!" Maybe if you're in a long-term committed relationship with someone you could bring up the topic politely, while remaining aware that their pubic hair grooming is their own business and you have no right to demand anything.

In the rare occasions where I've had the opportunity to find out men's opinions on pubic hair, they have often been enthusiastic. For instance, when I cammed, my clients universally preferred a hairy pussy. (As my ex-girlfriend used to joke, "the first day you cam you shave your pussy, six months in you start googling 'pubic hair thicker darker techniques.'") And of the men I know who have mentioned their opinions on pubic hair, most have been something along the lines of "I say grow that

shit like a jungle, give 'em something strong to hold onto, let it fly in the open wind" (although they do not generally agree that if it get too bushy you can trim).

Do I think the thinkpiece writers are wrong? Probably not! I suspect they're accurately reporting what the dating pool is like for them and their friends, but for some reason it's different. Perhaps men who hire camgirls are older and have more old-fashioned preferences, or hairy pussies are undersupplied in mainstream porn causing their aficionados to seek out handmade artisanal porn, or a hairy pussy makes the camgirl look normal and attainable and clients find this attractive. Many of my friends are queer; perhaps queers are different from heterosexuals, and this rubs off even on the straight men around them. Maybe I spend lots of time in sex-positive communities, and we've successfully created a norm of body positivity, which means that people feel it is rude to make negative comments on other people's bodies. Maybe it's something I haven't thought of.

Another example: a few months back, I was reading an argument about polyamory in which a monogamous man said that he knew that poly men didn't really have girlfriends, because their wives would shut down this whole poly thing the second they started spending $10,000 a year on their new girlfriend, as of course everyone does. My first reaction was to make fun of it: who spends $10,000 a year on a girlfriend? What the fuck are you buying her, a solid gold pony shoed with diamonds? I want someone to spend ten thousand dollars a year on me, that sounds great.

(Topher: "I think that probably involves a lot of nice dinners at fancy restaurants with expensive bottles of wine, and you have a phobia of alcohol." Me: "they can buy me tea instead! you know how much really good pu'erh I could get for $800 a month?")

To be clear: while there might be some extraordinarily wealthy poly person who spends $10,000 a year on their girlfriends, in my experience of poly communities this is not true. Typical dates include "taking a long walk", "getting a cup of coffee", "watching a TV show on Netflix", "being on Tumblr in the same room and showing each other cute cat gifs", and "taking care of a small child together". (Maybe that last one is just me.) If you get dinner, you can generally expect to split the bill, unless one person happens to be particularly poor or prone to forgetting their wallet, and the date is probably going to be at a $5 burrito place. In my experience, polyamory only starts getting pricey if you have to buy plane tickets to visit out-of-town partners or start letting all your partners stay in your house rent-free in the Bay Area.

But when I make fun of $10,000/year guy, I'm making the same error. I've generally only dated broke students, broke artists, and programmers, who while wealthy have a distinct tendency to drive old cars and refuse to wear any shirts not given to them for free. And even if I did go on a first date with someone who wanted to spend $10,000/year on me, I would wear sweatpants to the nice restaurant, not be able to find anything lacto vegetarian on the menu without custom-ordering some very depressing spaghetti with marinara sauce, and flinch away from the expensive wine as if it were a spider. At that point the question is just who rejects whom first.

Instead of assuming that the people I date are a random selection from the pool of All People Who Date Ever, I should assume that they're a biased sample: they're people I'm attracted to, who are attracted to me, and whom I even get a chance to meet and interact with at all. This is a pretty biased subset of humanity: no prizes for guessing why I don't typically date monolingual Swahili speakers.

And I'm unlikely to notice the other subsets of humanity even exist. While I can observe the existence of truck drivers, hockey fans, and other people far different from me, sex and romance are private, and I only get indirect evidence and self-report of what other people's sexual or romantic lives are like. It's particularly easy to assume that what it's like for me is what it's like for everyone-- just like it's easy for me to assume that everyone else zones out in the shower.

Therefore, I think it's a good practice to, when people make claims about dating or sex that seem ludicrous or bizarre to you, have as your first hypothesis that they are accurately describing some dating pool you are not in.

Beginners’ Meditation

I was about to meditate. There was a beginners’ meditation class three blocks away at the 14th Street Y. Seemed wrong to not check it out. Low risk, potential high reward. Non-zero story value. Might even learn something.

It was held in Room 403. Not in the gym. In the preschool. Not great.

I looked inside. I saw a circle of about twenty folding chairs. Sitting in most of the chairs were very old women, chatting. Whatever filter was operating was not subtle.

It said, you are not the target.

Welcome to Meditators Anonymous. My name is Zvi. Hi, Zvi.

I asked if I was in the right room, hoping I wasn’t. I was. Despite lowered expectations, I sat down.

They passed around a petition to keep the class going after its one month engagement. This was class two. Impressive customer loyalty. Was a there there after all?

Several minutes late, the circle was complete. Me. One instructor. One Asian man. One Amy Poehler lookalike. Twenty-two very senior female citizens.

We were told meditation was us getting to know ourselves. Accepting our own friend request on Facebook. I tried not to make too much of a face. I tried not to make too little of a face. Would have been a lie. People deserve honest feedback.

We went around, said our names and mentioned something that warmed our hearts this past week. To get to know each other. I noticed I was confused. What did that have to do with meditation?

Some mentioned small niceties. Most mentioned babies. They especially loved fathers with babies, doing heroic things like pushing strollers or carrying the baby.

Here’s to you, baby carrying father. Here’s to you.

My son Gideon had been born the previous Thursday. So I won.

I like victory. Big fan. Passed the test just like all the rest. But never really understood the reasons why I took it in the first place.

We were told to sit comfortably in our folding chairs. Not a beginner task.

We were told to sit up fully straight, in a relaxed position. Some people can do that.

We were told to keep our eyes open, aiming six feet ahead on the floor, to take our practice into the world. We were told to relax our shoulders, and put our hands on our thighs. Can do.

Focus on the breath, either at the belly, chest or nose. If we have thoughts, label that thinking and come back to the breath. She mentioned things we might be distracted by. They were distracting. I got briefly hungry. I came back to the breath.

That was it. The whole instruction. The meditation itself lasted maybe ten minutes. There were people fidgeting. The instructor eyed her watch. I had thoughts. Many were meta. Is that common? Perhaps common but not discussed. I came back to the breath. We finished.

The instructor went over ‘the practice’ again, asking us to name the steps. Like this was a classroom. It was. So, fair? Still infantilizing. Disrespectful of my time. This wasn’t complicated. Sit up straight, hands on thighs, look down six feet, focus on breath, bring focus back.

That was “the practice.” It had an exotic-sounding name.

Teacher said we keep our eyes open and sit in chairs so we don’t fall asleep. Must. Be. Nice. I envy those who fall asleep that easily. I am not enlightened.

I also wonder if that means that if you *don’t* fall asleep, you *should* lie down. Tempting!

Teacher took questions. Students pointed out focusing is hard. She agreed. They asked if it stops being hard. She said it doesn’t.

Teacher concluded with a call to ‘daily practice.’ If you had two minutes, that was all right. Two minutes will change your life!

How, exactly?

I know why I’m meditating. I’ve talked, theorized, modeled, gotten recommendations, read the review post, read the comments, read the other post, done better theorizing and modeling, and read the book up to where I couldn’t tell if it was talking nonsense anymore. Stage Four. Reading further would only distract. Beginner mind. So far so legit. Exceeds expectations. Tentatively recommended, could tell you why. Some day.

This class made no such attempt. No goals. No explanations. You *make friends with yourself*, on two minutes a day? No there there. Epic fail.

So why were *these old ladies* meditating? What kept them coming back, even signing a petition? Wasn’t point *you only need to come once?* Was this an excuse to share warm father and baby anecdotes? Why else would they want the class to continue? Why else were so many of them there so early?

Nobody Does the Thing They Are Supposedly Doing.

I hope that’s it. I wish them well. They seemed nice.

Bring Back the Sabbath

Epistemic Status: Several months of experimentation

Previously: Choices are Bad, Choices Are Really Bad, Complexity Is Bad, Play in Easy Mode, Play in Hard Mode, Out to Get You, Slack

For More Thoughts After: Sabbath Commentary

Alternate Take (Endorsed): Sabbath Hard and Go Home.

Slack is life. It is under attack. We must fight for it.

Choices Are Bad. Really Bad. We need a break.

Complexity Is Bad. We need a break.

Work is exhausting. We need a break.

Relaxation is hard. Our attempts fail or backfire.

The modern world is Out to Get You. We need a break.

We need time for ourselves. Time that is truly our own.

Without setting aside such time, that won't happen. Even when you take time, you'll be continuously *choosing* to take time, and… well, whoops.

Modern life made the problem worse, but the problem is ancient. The ancients had an answer.

We need rules. We need ritual.

We need the Sabbath.

Cabin in the Woods

The parallels of my and Ben Hoffman's Sabbath realizations are striking.

A few months ago, like Ben, I needed a break. My job puts me under constant pressure. My weekends weren't refreshing me. Like Ben, I experimented with camping. Like Ben, I had no spare battery, and left my phone off. I read The Great Transformation. I had meant to do that for weeks. I loved the world *leaving me alone.* Like Ben, I could relax, slow down, think.

I wasn't worried about things I *could* be doing – I couldn't do them.

Could I get this without the trip? Friends had started hosting Friday night dinners. What about the whole thing? What if we brought back the Sabbath?

Tradition makes rules easier to justify and explain, to others and yourself. These rules were time tested. I could take them and make them my own.

I thought about the components. Which made sense? What rules would let me cut the enemy, and relax?

Return of the Ritual

Rituals need clear beginnings and endings.

Sabbath begins with candles. One lights two candles, and recites a blessing.

For the evening meal, one says additional words and blessings, drinks wine, eats bread from one of two whole loafs and sits down to a proper meal with friends and family.

The candles are a signpost and deadline. Your week is complete and your work is done. There will be guests, so the apartment is ready. The ritual objects, and your needs for tomorrow, are secured. The meal is prepared. Time to feast and relax!

Slack is thus preserved in five ways.

This creates a time and place to see friends and family. Most want more social events, but coordination is hard and events are work. Now there's always Friday night.

They increase the value of improving your home. Every week you notice the little things that enrich meal, visit and home. They're Worth It, but easy to forget. Enhancing the little things enhances your life.

They prevent accumulation of personal-and-home-related work debt. A chaotic house is not restful. Postponed chores weigh on you. The deadline forces handling them in advance. Payoff is immediate.

By moving work *before* the deadline you are forced to *make time during the week.* You don't eat into Slack. If you can't find the time, this alerts you. Emergency!

They create visible failure as you approach hard bounds. When emergency arises, you sacrifice from the ritual. This signals emergency *before* life falls apart. You still have reserves. The ritual *is* Slack.

Sabbath ends with another candle. This prevents doing work until you *go through non-trivial motions*. You must do it on purpose.

Four Freedoms

We need restrictions that free us from the world. We need a new four freedoms.

We need *freedom from work.* Decide what counts as work to you. Don't do that. Anything done for money is automatically work. During the week, time is money. Today, do what *you* value.

We need *freedom from interruption.* Space to think. Cut off the outside world. Especially cut off anything continuously updating and all periodic rewards. There lie Skinner boxes. Much of the world is out to get you. Today it can wait. Friendly visitors are welcome, but ideally arranged in advance.

We need *freedom from choice.* Full freedom from choice requires a step *beyond* the traditional rules. In my version, even among permitted activities, only those explicitly

selected in advance are available – particular books, radio stations and so forth – plus things you feel intrinsic motivation to do. No lists. No browsing.

We need *freedom from stress*. Stressful conversations are *not allowed*. Doing work is *not allowed.* Making decisions is *not allowed.* Outside information is *not allowed*. If something was *still* going to stress you out and it was fixable, *fix it before the Sabbath.* Things can't change on their own, and you can't make them change. Why stress?

Sabbath Easy, Sabbath Easy, Sabbath Hard

Tension exists between *that which is most restful right now,* and *that which would be a stable set of rules.* There are *two* Easy Modes, representing each extreme.

One extreme is Orthodox Sabbath. This uses *a strict, fixed set of rules.* Pure deontology. You can't *carry objects* without special preparations. Many objects you can't even *touch.* This interferes a lot with relaxation, and *forces realignment of life to prevent that.* That can be good. There are even rules about violating the spirit of the rules – if you violate the spirit without breaking *even those rules,* that's almost encouraged. Restrictions allow maximization.

Another extreme is Reform Sabbath. This asks, *what would be most restful today?* This is utilitarian and uses causal decision theory. Sabbath is for rest, so if driving a car or making a call would be more restful, do that. You *could* break the rules. This destroys freedom from choice. Who respects such boundaries? You won't have urgency *before* the Sabbath. You can handle things later. Wouldn't that be *more restful?*

The Hard Mode approach asks, *what sustainable rule set best preserves long run Slack?* Taking stock and encouraging Slack-preserving outside the Sabbath are explicit goals. It uses logical decision theory. It creates personalized rules *you* can follow that work *for you,* but understands each divergence you select is expensive.

It asks *what would be in the spirit of the rules,* and modifying the rules to reflect that spirit. It views breaking current rules *during the Sabbath* with extreme skepticism, to reinforce following the rules. It modifies rules on Sunday.

In choice-related ways my current system is more restrictive than the Orthodox version. Mostly it is less restrictive, but becoming more restrictive over time. I currently allow Level 4 but everything there is on the chopping block. On Friday night I restrict to Level 2.

Hierarchy of the Shabbistic

There exists a hierarchy of shabbisticness. At one end are activities aligned with the goals of relaxing, recharging and unplugging. Sleep certainly qualifies. At the other are activities perfectly in conflict with those goals. Work done for money.

The hierarchy's details are different for different people. If you see something as work, it drains you. Move it down towards the unshabbistic. If you see something as invigorating, and have the spontaneous urge to do it *for intrinsic reasons,* move it up towards the shabbistic.

Then draw a hard line. Deciding whether to allow something is an impactful choice (itself banned) and a slippery slope. The golden rule of Sabbath is not breaking the rules. When in doubt, don't do the thing, then refine your rule on Sunday.

I encourage stricter rules for Friday night than Saturday. This enriches without being stifling.

This is my current hierarchy. Levels 1-2 I consider purely good, Levels 3 good, Level 4 questionable. Level 5 is bad, Level 6 very bad. *Level 7 is banned all week.*

1. Pure rest. Sleep. Rest. Walking. Intellectual discussion. Friendly discussion. Reading physical books and other physical objects. Meditation. Museums. Taking a bath.
2. Active rest. Sex. Flirting. Running. Swimming. Playing sports. Arguments for low stakes. Board and card games with no stakes. Puzzles. Building models. Taking a shower. Eating. Watching sports in person. Light switches.
3. Consumptive rest. Riding elevators. Radio with one station. Listening to music. Food preparation without lighting a fire. Window shopping. Kindle and other e-books.
4. Potentially toxic actions. Writing for yourself. Taking notes. Practicing and training personal skills that are not work or work related. Working out. Computer games. Pre-selected television. Phone calls and texts for physical coordination purposes. Riding in cars and trains (without payment).
5. Violations of compactness. Phone calls and texts not for same-day logistical coordination. All other use of smartphones. Making impactful decisions. Planning. Flipping stations on television or radio. Browsing the internet. Browsing a giant music collection. All long lists, especially lists of choices. Checking anything that continuously updates. Lighting a fire. Stressful topics of conversation.
6. Work and outside demands. Exchange of money. Doing business. Anything that earns money or creates commercial value. Negotiations. All continuous updates. Email.
7. Considered harmful. All timed and daily rewards. Micro-transactions. Social media.

The Rules Simplified

Start here. Adjust as needed.

Light candles before sundown Friday to begin.

No outside inputs except in person.

No choices impacting post-Sabbath.

Light and extinguish no fires. Do no work or business. Spend no money.

Only preselected and spontaneously motivated actions are allowed. No browsing. No lists.

Light another candle after sundown Saturday to end.

State of Emergency

I brought back Sabbath for Slack and relaxation. Ben brought it back as an alarm system, for when life was out of control. Sabbath shows when you are not okay, and provides method and incentive to get back to okay.

This Saturday I did full Orthodox Sabbath (minus prayer), and also fasted, as an experiment. I won't do this every week or even month, but it had important alarm value.

Ben's post is excellent. Read the whole thing. I'll finish with two key passages from it.

Key motivation:

> You would not want to do this sort of thing all the time. But it might make sense to do periodically – perhaps once a week – as a stopgap measure to combat attention drift. If powerful and pervasive cultural forces are out to get you, you ought to check in from time to time with yourself, and other people with whom you have local, high-quality relationships, to give yourself a chance to notice whether you have gotten got for too much.

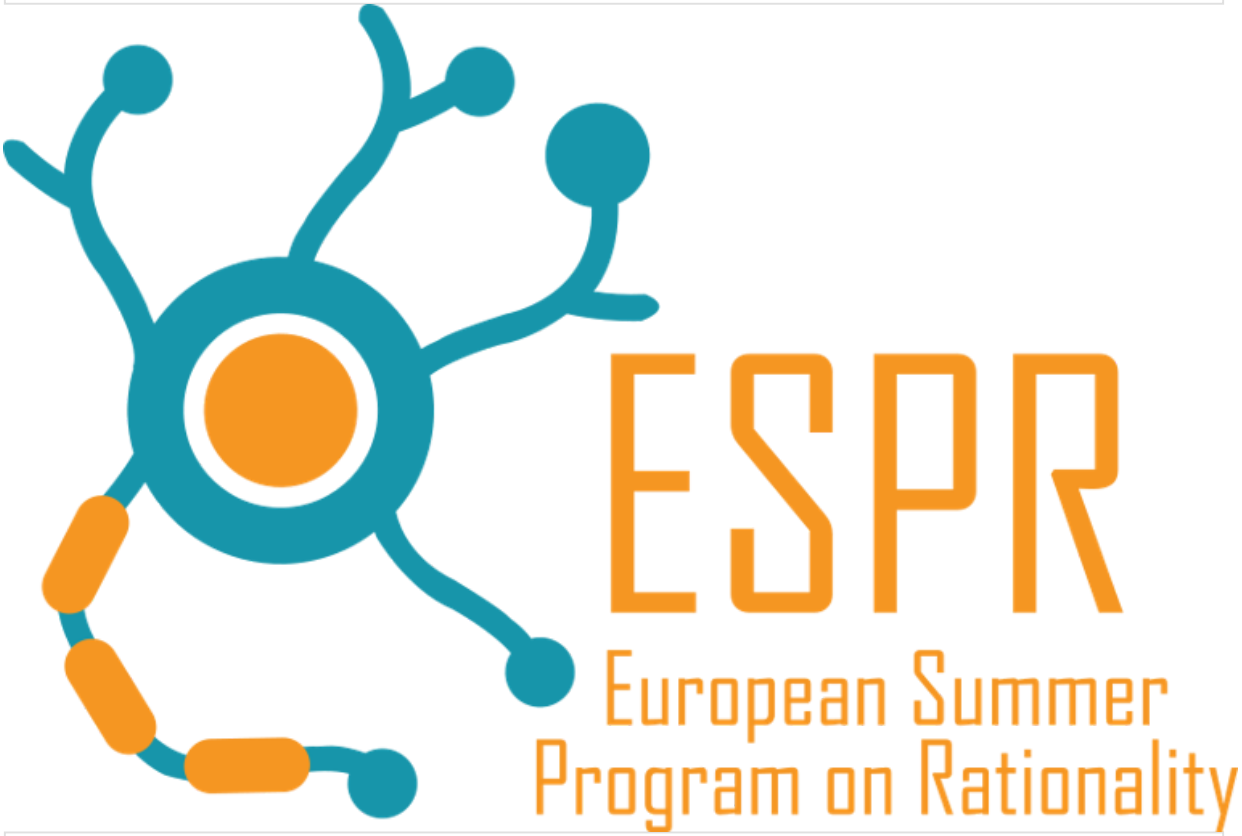
His conclusion is important and worth quoting in full:

> One more useful attribute of the Jewish Sabbath is the extent to which its rigid rules generate friction in emergency situations. If your community center is not within walking distance, if there is not enough slack in your schedule to prep things a day in advance, or you are too poor to go a day without work, or too locally isolated to last a day without broadcast entertainment, then *things are not okay*.
>
> In our commercialized society, there will be many opportunities to purchase palliatives, and these palliatives are often worth purchasing. If living close to your place of employment would be ruinously expensive, you drive or take public transit. If you don't have time to feed yourself, you can buy some fast food. If you're not up for talking with a friend in person, or don't have the time, there's Facebook. But this is palliative care for a chronic problem.
>
> In Jewish law, it is permissible to break the Sabbath in an emergency situation, when lives are at stake. If something like the Orthodox Sabbath seems impossibly hard, or if you try to keep it but end up breaking it every week – as my Reform Jewish family did – then you should consider that perhaps, despite the propaganda of the palliatives, *you are in a permanent state of emergency*. This is not okay. You are not doing okay.
>
> So, how are you?

Unofficial ESPR Post-mortem

[Disclaimer: This post reflects my, i.e. Owen Shen's, personal opinions. It does NOT reflect CFAR or ESPR's opinions, and should NOT be taken as an ESPR-endorsed communication.]

[An overview of some of the camp and project dynamics of ESPR 2017. I look at diversity causing certain problems, difficulties evaluating participants, ways to improve camp experience, and organizational difficulties.]



Introduction:

In August of 2017, I went off to London for the European Summer Program on Rationality (ESPR). This was supposed to be the place where all our efforts over the past 8 months came together, and we made The Best Summer Camp Ever.

ESPR is a two week summer program for highly talented students 16-19 across the world. We had talks/classes on machine learning, cognitive psychology, effective altruism, cryptocurrencies, and salsa dancing. (Note that everything except for the dancing may not be representative.)

I was a participant of ESPR 2016; last year, it was called EuroSPARC. The camp name and spirit is modeled off the US-based SPARC camp. For ESPR 2017, I ended up being responsible for (part of) admissions, communications, assorted pre-camp logistics, and camp counselor duties.

This post is structured as a sort of free-form analysis, where I examine a section of ESPR, look at how it worked in practice, and then move on to another section. There are some common threads between these sections, but there's not really a central thesis; there's just a series of lessons that I learned.

Also, like most post-mortems, any analysis is undoubtedly going to focus on the Bad and the Sub-optimal. To provide some early counterbalance to the ensuing dive into things which ended up not working, here's a quick overview of what went *well:*

- 72% of our participants rated us 9/10 or 10/10 when asked how glad they were that they attended.
- Unexpected setbacks which could have critically ended ESPR were miraculously solved by heroic staff in the nick of time.
- Multiple students took initiative during camp and taught their own classes, grabbed us a guest speaker, and other exciting things. (See the section on **Improving Camp Experience** for more details.)

Thus, ESPR had many encouraging signs that things went well.

Of course, a string of things also went wrong. Volunteers left, interpersonal conflicts were raised, and it's not clear that our participants got much in the way of concrete takeaways.

I hope that looking at all these different pieces of the project can shed some light on why some of the bad things happened, and how we can do better for similar projects in the future.

tldr;

Here's an overview of the points that are covered below in the following paragraphs:

1. This year's diverse curriculum arguably decreased the value of takeaways for the participants.
2. A vague mission statement meant that different ESPR staff had sometimes conflicting goals which weren't well resolved.
3. Evaluating what participants got out of ESPR is a very hard task because there are lots of relevant factors.
4. One strong way to improve the ESPR experience for participants is to bring in more opportunities for independent projects and ownership of learning.

5. Unclear role and responsibility specification led to negative incentives for people to take initiative on tasks.

Initial Expectations and Values:

I'm going to be evaluating different parts of ESPR, so it seems crucial to give an outline for which things I cared about and found relevant.

First and foremost, I had originally wanted ESPR to be able to churn out people who'd be motivated to work on problems in the rationality space and effective altruism space.

(This has now actually changed; see the **Evaluating Participant Takeaways** section for more details.)

A big part of this was that, as an ESPR 2016 alumnus, I had found the camp to be important in roping me in deeper to these ideas, which I consider valuable.

The actual effects of my slant on camp was tempered by others, who had different views, giving pushback. (More on this idea in the **Diversity and Dilution** section.)

But I just want to set the framing that, compared to other staff on the ESPR team, I was probably one of the ones who leaned the most in the direction of rationality / EA.

Diversity and Dilution:

As I mentioned above, there was some strong pushback by others to make ESPR less directly about *just* rationality and effective altruism. The end result meant that the camp ended up being quite different than just "CFAR, but for super smart high school students".

And I don't think this was bad. In the end, I agreed with much of the concerns raised with the pushback, and I've also updated my thoughts on what exactly ESPR should be trying to give students. (More on this in the next section **Evaluating Participant Takeaways**.)

Still, something that I don't think I fully appreciated at the time was that diversity has costs because the camp's focus is a zero-sum resource. Any time you introduce a wider range of topics to cover, it's conceptually less clear for the students what the "important things" are.

(I'm mainly talking about curriculum here; it's not clear that it generalizes to other activities, given that "classes" and "not-class-things" already have a clear conceptual boundary.)

A phrase that was often repeated at camp was, "I have no idea how I'm going to explain ESPR to my friends and family when I get back home. There's been so much that happened."

This sentiment, to me, seems to maybe be a bad sign. There are definite reasons to want a camp to be set up like an experience that travels across many domains in a way that makes it hard to easily articulate, but this also makes it harder to track what happened.

For me, I think it boils down to the question of whether or not the takeaways are stronger when people need to put in effort to explicate them (as would be the case if the initial feel was “hard to explain”) or when they are more clearly chunked (as would be the case if clearer demarcations and perhaps less subjects were covered).

I think one of the stronger benefits for having a more diverse curriculum is that students are able to pick and choose the takeaways they want. If you cover lots of things, then it’s more likely you’ll hit upon something that resonates with them.

Yet, while this does seem like a good thing that could have happened, we also didn’t stress that this sort “take what you need” was the appropriate attitude during camp. Overall, I think the end result was that a lack of a good structure for students to organize what they learned.

Another related barrier for takeaways was ESPR’s focus on bringing together students from a mix of countries; ESPR 2017 had participants from a greater range of countries than ESPR 2016. This led to more difficulties in communication than last year, and I think we didn’t account for this heavily enough in the initial planning.

Cultural expectations within the classroom (EX: how the “teacher” and student dynamic played out) and other social norms felt like they contributed to the wide spread of attitudes and ideas, at least initially.

Of course, one thing I also heard a lot of people say at ESPR was, “One of the best things here is meeting people from lots of different cultures and seeing how things are in other places.”

The obvious answer to the above is “Find the sweet spot between having a mix of cultures and cultural commonalities.” We couldn’t just optimize directly, though, as lots of staff had lots of opinions which differed.

In an effort to bring in people from lots of places, we launched ESPR with a fairly vague mission statement. This meant that we had people under our banner who held conflicting implicit goals, despite the fact that, outwardly, we all supported the overt, non-controversial mission statement.

For example, I very much wanted our participants to get more into effective altruism and rationality. Other people wanted ESPR to be a more playful experience, where students could explore a variety new ideas (and not be shoved face-first into certain ones).

One problem this led to was happening discussions about what sorts of values we should have *during* camp, rather than *before*. I think that, had we been more explicit about our goals upfront, we could have worked around some of these problems.

Specifically, we could have brokered compromises early on to ensure that people knew what they were getting in return, and we knew what everyone actually wanted (and that a discussion was happening on how they could get it).

Failing to account for this when bringing on a diverse group of people meant that we had to regress to the broad mission statement (once again diluting things), and were less able to satisfy many of the staff’s specific preferences, or come to acceptable agreements for both sides.

Evaluating Participant Takeaways:

The whole reason we put in effort to make ESPR 2017 happen was because of the participants. Though everyone on the staff might have had diverging agendas, we all wanted *something* to happen to the participants as a result.

Evaluating ESPR's impact, however, is a difficult process. Group dynamics are complicated to model, there are lots of confounding factors, and getting clear indication of participant growth (let alone counterfactual growth!) is hard.

With all that in mind, I'll try to dive a little into what my thoughts are on exactly what the effects were.

First off, I think it seems reasonable that the three largest factors in influencing the experience of camp for the participants were:

1. The curriculum (EX: What topics were covered.)
2. Additional activities which encourage ownership (EX: Getting students to do their own independent projects.)
3. The admissions process (EX: Who we took in.)

From a combination of survey data and first-hand impressions, I think that about 67% (~20) of the participants came away with *some* kind of takeaway. In terms of EA/rationality orienting, though, it seems that about 17% of the students (~6) got a lot out of ESPR.

It's not clear, though, that these are the base rates of "what percentage of people who go to ESPR come out awesome?" There are several confounding factors here.

One of the biggest ones is that most of the people who seemed to pan out very well (i.e. were good along the EA/rationalist axes) were also people who we knew were already exposed to these ideas prior to camp. This also seemed to be true for last year, and I count myself as one of them.

So, back to the overall goal of ESPR, there's a crucial question of "rope newcomers into the community" versus "marshall existing young community members" which I think we partially sidestepped early on.

Part of this had to do with ESPR's focus on high achieving participants, for example those who had succeeded in national and international math competitions. Here's one version of the story for why we focusing on these students from a consequentialist position might make sense:

"Most of the important discoveries are going to be done by people who are at the top of their field. Thus, making sure that the future is going to turn out well means making sure the smartest people in 20 years have a good set of ethics. Thus, we should focus on those people."

And while I don't think any explicit form of the above reasoning took root in our decision process, I definitely think it was present in some form.

Now, though, I think that trading off additional technical ability for pre-exposure to EA/rationalist is generally a good idea if the thing you care about is impact, and the actual benefits work out in its favor. In other words, we should weight attitude more than we currently do in the attitude-aptitude trade-off.

This is a direct result of my experience at ESPR 2016 / 2017.

(Note that this may still be too small of a sample for my opinion here to mean very much.)

I've also changed my mind on what I want participants to get out of ESPR:

As a result of conversations with others who gave pushback on explicitly pushing for the rationality and effective altruist angle, I think I've now pivoted to thinking that the point of ESPR is to get more people thinking altruistically, with effectiveness and rationality as merely nice-to-haves.

What I sort of mean by that is a set of questions that sort of goes like, "Do I expect this person to be doing exciting work that will, on net, be beneficial to other people in the future? Do I expect them to care about helping others? Do I see them interacting with others who are engaging in humanitarian projects?"

(I know the above criterion is still vague, and it merely sort of maps onto my gut impressions / checking in with my internal anticipations to see what I expect. It also feels like the best I've got for now.)

Anyway, this is a noticeable shift from my original viewpoint of "Let's get everyone super into effective altruism and rationality!"

While part of this is arguably due to holding more realistic expectations about base rates for memetic adoption (read: the onset of cynicism), there is another part of me that's genuinely unsure about how stable or correct the EA / rationalist frameworks are. It just seems like a good idea to have people who share similar values exploring a different space, even if my values were deeply EA-aligned.

Knowing all this now will probably make some things clearer if / when I start to consider applicants for ESPR 2018.

But evaluation is still difficult. Above, the 67% and 17% estimates were only for benefit the *participant* received. Yet, there's other ways a participant could be a good pick, even if they themselves didn't get that much out of ESPR.

Consider the following factors, which would all seem to indicate that, from a consequentialist viewpoint, it was "good" for ESPR to take on a certain student:

1. Degree to which the student contributed positively to the camp environment (EX: Alice was responsible for making camp better for other students.)
   Relative difficulty to measure: 1/4

2. Degree to which the student received concrete takeaways from the camp (EX: Bob came away with a new outlook on life.)
   Relative difficulty to measure: 2/4

3. Degree to which a student's received benefit was counterfactually good. (EX: Carol had a very positive learning experience at ESPR, relative to all her other options for the summer.)
   Relative difficulty to measure: 3/4

4. Degree to which the student will go on to have a positive impact on the world (EX: Evan leaves ESPR with grand ideas and creates a startup designed at

providing free global WiFi.)
Relative difficulty to measure: 4/4

Looking at both the ESPR 2016 and 2017 cohorts, my impression is that, while some clustering in the above factors occurs, it's also really hard to predict this type of stuff ahead of time, as well as measure this post-camp.

In fact, I'd find it plausible that trying to subtly control admissions for the above factors is largely useless. During admissions, we made certain bets on which students might fit the above factors, EX: which ones we expected to contribute a lot to the overall atmosphere.

Some of those bets panned out, but some of them didn't.

As a result, while I think it's true that admissions has a strong overall effect on the camp, we can't escape base rates. It seems likely that, at the end of the day, we'd also have seen roughly the same 67% and 17% percentages.

A good part of my intuition for this comes from the fact that "probability that a student takes well to crazy-sounding ideas" (which is what at least 30% of what ESPR is, I claim) seems to be distributed about the same, largely independent of who we pick.

This overall seems to be a point *against* admissions being important.

The outside view says that the communication channels we used to advertise ESPR actually did a lot of the heavy lifting of filtering for capable candidate (which might in itself be a positive of focusing on highly talented students, contra my "attitude over aptitude" stance).

Still, my internal estimate says that, even if the 20th best candidate wouldn't have differed greatly from the 50th best candidate, *surely some type of filtering by the interviewer (i.e. me) was happening was relevant to why they were even in the final pool, right?*

And perhaps the right response here is, "Owen, you're being fooled by noise!"

("I can't hear you over the sound of my self-righteousness!")

Actually, though, I think the right answer is that interviewing is pretty good at avoiding Type 2 errors: Most of the people we rejected after interviewing were (I subjectively claim), probably not that great.

However, the fact that our bets didn't all pay off seems to indicate that we're going to get false positives, and this is something we can't strongly remove.

Improving The Camp Experience:

While it's debatable, then, what effect admissions selection has on the camp experience (which is why I listed it last), it seems clearer that what actually happened at camp had a large effect on the participants.

There is a story, albeit one that I don't endorse, where the majority of the value of ESPR comes from simply bringing smart people together. As this model goes, the actual classes themselves aren't as important as the social aspects dominate, and the social aspects are responsible for most of the good things ESPR offers.

So, within the context of class curriculum, camp events, and evaluating their impact, here are my two thoughts:

1. Most of the benefit to participants from ESPR can be traced back to just a few (albeit, different) events.

2. Providing opportunities for ownership of ideas / learning is one of the most important pedagogical techniques, and we underutilized it.

The first point is basically a consequence of how humans compare things relatively.

Even if all the events we provided at ESPR are super-duper great, there's going to be implicit ranking and comparison happening in the students' heads. While I think this means we could strategically aim for certain events to be "The Important Ones", I also think we can't get too good estimates on which events will land well (with a few exceptions).

This ignorance doesn't mean, though, that we're allowed to slack off on quality for some activities. Rather, it feels like having every activity be Awesome is a *prerequisite* for those relatively life-changing opportunities to crop up in the first place.

The analogy here is how you have lots of thoughts everyday, but it's likely only a few of them are insightful. But you can't get them in the first place by trying to think just a few thoughts; if you aren't always thinking, they just won't come.

The second point is about giving students more opportunities to exert their own Do Thing muscles. Overall, I feel like we still emphasized learning over doing.

I think it's clearly Good that ESPR made time for students to assert their own abilities and own up to crazy ideas. This manifested in letting students teach afternoon classes, sending them off into the world, and assigning one another Quests to accomplish.

I think two of the most successful activities in this spirit that we ran at ESPR were Lightning Talks and Social Engineering. During Lightning Talks, every student was encouraged to give a talk on a topic (any topic! English tea! Wales! Freud!) for about 2 minutes. Social Engineering was, in a nutshell, when we sent students off into the streets of London to convince companies to do nice things for us.

There seems to be something that's just Good© about letting students do projects / schemes, and I think we didn't mine this area nearly enough for all the good things that could have popped out.

And for what we did do, the results we got were always interesting and, surprisingly often, impressive. (Guest speakers! Deep dives into humor analysis! Free lunches!)

Camp Internals:

When it came to the actual interactions between the ESPR staff, I think the biggest problem we faced was the unclear designation of roles and a lack of specificity in what the roles were supposed to do.

Like most things in life, no one was being actively stupid in not spotting this beforehand. It just happened that there was a chain of understandable events which led to this being the state of affairs we found ourselves in.

But the end result was still such that, by the time ESPR started, there still wasn't mutual understanding or agreement on which tasks each role (student, counselor, instructor, ops, etc.) was responsible for. I think this ended up being harmful to some staff, me included, as well as overall camp operations.

First off, a lack of clear duties for each role meant that there was a lot of inter-role variation. Staff who were supposedly doing the same thing in name would end up having quite different tasks in actuality. This is similar to the issue with divergent goals all flocking under the general mission statement banner.

I think this ended up breeding bad sentiment (at least for me) when the benefits associated with each role, which *also* weren't well-specified beforehand, ended up poorly calibrated, i.e. incommensurate, with the amount of effort different staff were putting in.

The obvious thing to do, then, is to once again try to be upfront with each volunteer about what they were hoping to give in and get out of the project. And I think we sort of did this.

The problem here was that these expectations changed over the course of the project, as some people did more than we initially brought them on for. People's actual roles stretched and shortened throughout the extent of the project, and motivation often rose or dropped for individual staff. And as the situation changed, earlier promises couldn't always be fulfilled, or people asked for more.

Related to this is the fact that different criteria are used to evaluate competence for each of the roles, and this competence doesn't necessarily translate from domain to domain.

This meant that ESPR was placed in a dynamic where some people who went the extra mile weren't always compensated the way they wanted to be because they would perform poorly at those roles. For example, an outstanding instructor might have wanted to also try her hand at operations work, but we ended up saying no to the request because she likely wouldn't have been very useful in ops.

There was a definite trade-off, then, satisfying between people's preferences and what would be "best" for the camp overall.

Once again, I think being more upfront about all this and spending more time to find compensatory measures would have been very useful in curbing some of these issues.

Even in strictly volunteer-only roles, I claim that people always have tacit expectations for what their give / get relationship with the project "should" look like, and our inattention in this area meant additional burdens were borne out unequally by certain ESPR staff.

Lastly, I think there is often a type of pressure for groups to put tasks up for grabs, in an Everyone Does Everything sort of way. There's something that feels virtuous and democratic and "nice" about this setup, as it sends a signal that roles are irrelevant and everyone is equal / pitches in.

I want to push against this intuition.

Putting tasks up for grabs in a democratic way negatively incentivizes first-movers who take the initiative. It's a tragedy of the commons, where everyone is individually better off if they wait for another person to shoulder yet another task.

But it’s not just this dynamic; I claim it’s worse than that.

Tasks you accumulate snowball, *further* punishing people who take the initiative. This dynamic pulls people into deeper roles which, if the responsibility / reward structure isn’t laid out well, breeds burnout and bad feelings.

Here’s an example that actually happened:

Right before ESPR started, we needed to figure out transportation for the students. As I had some spare time, I ended up doing the preliminary logistics work of cataloging all the info we had into a spreadsheet. But this also meant that for the next action we had to take on transportation, I was also the go-to person because I “had done it last time”.

The cycle here is pretty vicious—with each additional task that I did to further our transportation goals, it also became less likely that I’d get additional help from other people because the *required background knowledge increased*.

EX: “I could also rope in Alice to help with this task, but then I’d have to tell them about X, Y, and Z. Oh well; I guess it’s more efficient for the group if I just do it again then.”

The obvious answer here is perhaps to fight against the sunk-cost response that appears in the above example because roping in an additional person in the know will actually *save* time down the road. But, really, what I think we need is just a better way of letting people take ownership of certain tasks. Ambiguity in delegation is a recipe for inaction.

I think prior to ESPR, I might have also tried to champion an organizational structure along the lines of Everyone Does Everything.

I no longer endorse this, and, as a result of my experience, lean towards a much more top-down / legible and explicit approach when it comes to distributing tasks in an organization or project.

Personal Conclusion:

So that's a lot about how ESPR The Project functioned.

For Owen The ESPR Staff (i.e. me, personally), my experience was largely very positive and it was paired with, as I've perhaps alluded to, several disappointments.

I felt like we missed some of the cohesiveness and "spirit" that was present in ESPR 2016, but this was partially remedied by the inclusion of new bright spots, like seeing last year's participants step up as counselors or this year's participants teaching their own classes.

Overall, I took on *significantly* more responsibility than I originally signed up for, and diving deep into the thick of things was highly instructive. Most of the insights, though, seem to be in the form of gut-level expectations, and this post-mortem has been an attempt to tease out some of those models and intuitions.

For the rationality community at large, I'm unsure what the takeaways are. Student outreach is always tricky, and there's much more to be said about things like setting camp culture and how different students interact with the entire rationality memeplex.

As a case study of how several people come together to form an organization and then execute on a major project, holding goals roughly aligned with the community at large, I think we can serve as a useful example.

There's a definite change in the quality of a model when it comes to knowing something vs experiencing it, but it's my hope that this can contribute to the ongoing discourse on group rationality as well as pedagogy.

React and Respond

Some may find it annoying, but one of my favorite things about the English language is that it's full of synonyms that point to roughly the same place in concept space but with each word being a distinct vector carrying nuanced differences in meaning. We see this with "react" and "respond", which both roughly have a meaning of "do something because something else happened", but each of which carries its own implications about the how the doing is done.

"Respond" is the older word so we'll start with it. It comes from Latin through French and is constructed from the roots "re-" and "spondere". "Re-" serves like "back" in its prepositional sense in English as in "go back there" or "do it back to him" and is synonymous with "in return". "Spondere" means "to pledge", but over time "respond" morphed to mean something less like "make a pledge back" and more like "answer back" or "answer in return". The result is our modern word "respond" that connotes a deliberate action made in return to another one, carrying some but not all of the original weight of "pledge" with it.

"React" is newer. It first appears in the 1640s in the sense of a physical reaction from "re-" plus "act" from Latin "actus" which itself draws from Proto-Indo-European "*ag-" which covers the notions of driving, drawing out or forth, and moving. So we can say "react" is to "move in return" to another movement, and it carries that sense through today in a generalized way as both the word we use to talk about what chemicals do when they are "moved" to react and the immediate return actions people make when acted upon.

Together "react" and "respond" create a dimension along we can describe what is done in return. Specifically, we more call those things "reactions" that are done quickly, automatically, or without deliberation and more call those things "responses" that are done carefully, voluntarily, or with deliberation. To give an example, suppose you go to see a movie about factory farming. While watching it you *react* with disgust, horror, indifference, etc. and may even post about your feelings on social media. After you think about what you saw, though, you may *respond* to it by changing your behavior or talking about why you think it was distasteful content to show. In this way "react" and "respond" roughly correspond to System 1 and System 2 generating action in return to experiences, respectively.

But that's just a first gloss because I think the S1/S2 distinction is confounded and something more is going. When "react" and "respond" are construed as I have construed them above, most people (especially the sort of people who read LW) would say they would prefer to "respond" to "react". That is, they would prefer to think before choosing an action rather than autonomically acting. Yet, no matter what we do, we will have some autonomic reaction, even if it is the "null" reaction of disregard, so even if we want to wait to respond we must react in some way first. Thus we get at one of the core challenges in rationality: how to react so that we may respond.

Since we already have mathematical models that mostly tell us how to respond if we have enough resources, even if we rarely or never have access to those resources and so must respond in suboptimal ways, it's tempting to create a space where reactions don't count and only responses matter. But to me that seems a dream because for the time being we are all humans and we all react to others reactions, so even if we try to suppress reactions we cannot escape them. Thus it seems we have no choice but to

react, and if we must react, it seems to me we might as well react in ways that accord with how we would respond.

So as much as rationality feels like a System 2 or far construal mode activity, it ultimately makes demands on System 1 and near construal mode thinking. As such we must engage with gnosis or personal experience of rationality if we want our reactions to align with our responses and to do the things that actually win rather than merely try. It seems only in this way that we may learn to react so that we may respond.

Avoiding Selection Bias

[This post has been renamed from "Desilencing", pending changing what I call the action.]

edit 2019-06-25: *this post is tone-deaf about ways people who experience the more common and dramatically stronger silencing forces of prejudice would see it. The insight here appears to me to be valid as an incremental change in an environment with low but nonzero hostility; it's not as immediately relevant when a very large change is needed. I have changed my vote on this post to a downvote.*

I often find that I filter my urges to give feedback, especially negative feedback, in public.

For example, when downvoting someone, I often feel an urge to say why. But then I hesitate because I worry that they will feel insulted, and attack me for my trouble of explaining myself.

This fear is not unfounded. sometimes when I say why, people do in fact challenge it.

But if I was on a discussion board with a bunch of slightly different myselves, and I never gave the other mes feedback, I would never get any feedback from them.

So, some semi-random fraction of the time, I say the thing anyway, in a short message with little overhead for me. I'm taking some risk, because then I say things that might get me in a fight. But people get more detailed feedback, instead of simply being ghosted or downvoted away because I'm scared of the fact that it's unsafe for me to be straight with them.

So when I say I'm "de-silencing" myself - this is what I mean.

I call it "de-silencing" because I do it to break the attractors that silencing forces on me create. Some non-negligible portion of the time, those forces do specifically intend to silence people. And this technique would not work if someone was specifically out to get me.

This post is itself a de-silencing post: I'm not putting as much effort into it as I think would be necessary to ensure it gets a good reception.

Identities are [Subconscious] Strategies

We all have identities. Arguably, any statement of the form “I am a __” is an identity. Of course, we usually reserve the term for the statements which feel especially core to us in describing and predicting ourselves, and in expressing our values and aspirations. Such identities may have their benefits, but they also come with a number of perils. In particular, we are prone to a) become distressed by any perceived threat to an identity, and b) become utterly inflexible around shifting, modifying, or discarding identities. These effects can be detrimental to both personal wellbeing and goal attainment. Sanity, however, can be regained if we recognize that our identities do not exist unto themselves, and are instead [often subconscious] strategies towards the attainment of specific goals and values.

Identities are Strategies towards Goals

Consider a person who prides themselves on their identity as a writer: “I am a writer.” This identity is precious because there is an implicit statement of the form “I am a writer[, and therefore I will have a job, income, status, friends, lovers, and my life will be good].” The implicit statement is the goal to be obtained and the explicit identity is the strategy for achieving that goal. The value of the identity derives from the goal is supports.

I describe these plans as subconscious because more often than not they are not articulated. Many people have an identity around being intelligent, but I expect that if you ask them why this important, they will need a few moments to generate their answer. I also expect that in many cases the belief in the goodness of an identity is absorbed from society and it is social drives which motivate it for an individual. In that case, the full identity statement might go “I am a __ [and therefore society will approve of me]” whether or not an individual would admit it. In the most general case, it’s “I am a __ [and therefore ***goodness***].”

Threats to the Identity are Threats to the Goal

Given that an identity is a strategy for achieving a goal, any threat to the identity is a threat to the goal. The degree of threat perceived is proportional to the importance of the goal and to the extent that the identity is sole strategy for achieving the goal. If someone believes that being a writer is their sole avenue for having a good and fulfilling life, they are going to get upset when that identity is challenged. This holds even if person does not consciously recognize that their identity is part of a plan. It is enough that some part of their mind, S1 or whatever, has firmly stamped “being a writer” as critical for having a good life.

Consider, though, someone who has identities both around being a writer and around being a musician. Suppose that this person has achieved considerable fame and fortune as a musician and resultantly already has wealth, friends, lovers, etc. by dint of this identity alone. I predict that this person will be less bothered by challenges to

writing ability than the person who staking themselves on being a writer. If the writer-only has their manuscript rejected, it will be devastating, whereas for the writer-musician, it will be a mere disappointment.

Protect the Goal and the Identity Can Be Free

If threats to identity are really about threats to goal-attainment, then the key to working with identities becomes a) surfacing the hidden goals and, b) ensuring there is security around attaining those goals. Tell the child that they're not cut out to be writer and they'll tantrum, but tell them they're not cut to be a writer yet have phenomenal painting skills, and they might just listen. Substitute one less viable plan for a new and better one. Other variations include exposing that the goal in fact has already been attained, as in the case of the writer-musician above, or recognizing that the identity in fact is going to be an ineffective plan regardless, e.g. giving up on being a goth because you realize that no one thinks goths are cool anyway.

Times I Shifted Identities

Recognizing my true goals and whether or not they are likely to be attained has allowed me to become flexible around, or even completely discard, identities which I have had for years or even decades. When I decided that I'd become an engineer in 2008, I formed a strong identity around being a "technical person." This was pure goodness to. Yet after deciding that my top priority was the long-term flourishing of civilization over building cool things right now, I realized that I needed to consider whether pursuing a technical path was really the best thing I could do. This was always an immensely difficult thing to do. I felt that definitely one day I'd move out of doing tech work, but not so soon! Recently though, I did make such a switch. By recognizing that a new path would maximize my values far more than the old, I transitioned from a Data Scientist role to a Product Manager role with little pain. I was giving up on spending my time of technical problems, but this path was truer to my real values and unquestionably was the one I must take.

Another identity triggered the line of thinking which led to this post. During a religious youth, I absorbed that anything bodily is crass and bestial and is only the mental and intellectual which are dignified. Though I left my youth far behind, this message stuck: the biological is undignified, shameful, bestial; only the mental and intellectual are good. For years, eating, drinking, sex, physical pleasure, etc. have felt embarrassing to one degree or another. My mind insisted this attitude was a core value of mine -- terminal, unquestionable. Sure, it suspiciously looked like the message I'd been taught as a kid, but it just felt so damn certain. Conceivably it was, but it still meant I lived in a state of conflict. After all I am still human with largely normal human biological drives and needs and I was doing those things even while ashamed. One way or another, I needed to get some kind of resolution around this identity.

To that end, at a recent CFAR workshop I began earnestly questioning what was going on in my head being open to any outcome. To my surprise, the exercise made me feel threatened in the exact same way I feel when socially threatened. This was odd -- if the anti-biology/mental-only thing was a core value, then I shouldn't feel socially threatened. In which case it couldn't be a core value. No, it was a deeply cached alief that it was shameful to be biological and that others would judge me if ever I unabashedly sought biological pleasure. Of course, that's absurd! No one I associate with feels that at all. People revel in pleasure of all kind! When I thoroughly exposed

the anti-biology identity to the evidence by just thinking about it for a few minutes, the identity relented and gave itself up. I was suspicious for some time, but I now enjoy all forms of pleasure without guilt or shame.

I remain surprised, and yet exposing and protecting underlying goals continues to work. A final example: for many years I've had a palpable desire to be wealthy -- an identity which has long interfered with my ability to consider non-profit work. I felt that I could only move into non-profit work, my eventual plan, once I'd saved up sufficient wealth. When I recognized that the identity of "wealthy person" was really just about status, and that already I'm pretty content with my status, this identity evaporated too. The goal it served had already been attained, the identity just hadn't got the memo yet.

To close, the questions to be asked are: What are you identities? Which goal does each of them serve? And do they serve them well?