Best of LessWrong: September 2014

Best of LessWrong: September 2014

[meta] New LW moderator: Viliam_Bur

Some time back, I wrote that I was unwilling to continue with investigations into mass downvoting, and asked people for suggestions on how to deal with them from now on. The top-voted proposal in that thread suggested making Viliam_Bur into a moderator, and Viliam gracefully accepted the nomination. So I have given him moderator privileges and also put him in contact with jackk, who provided me with the information necessary to deal with the previous cases. Future requests about mass downvote investigations should be directed to Viliam.

Thanks a lot for agreeing to take up this responsibility, Viliam! It's not an easy one, but I'm very grateful that you're willing to do it. Please post a comment here so that we can reward you with some extra upvotes. :)

Simulate and Defer To More Rational Selves

I sometimes let imaginary versions of myself make decisions for me.

I first started doing this after a friend told me (something along the lines of) this story. When they first became executive director of their organization, they suddenly had many more decisions to deal with per day than ever before. "Should we hire this person?" "Should I go buy more coffee for the coffee machine, or wait for someone else deal with it?" "How many participants should attend our first event?" "When can I schedule time to plan the fund drive?"

I'm making up these examples myself, but I'm sure you, too, can imagine how leading a brand new organization might involve a constant assault on the parts of your brain responsible for making decisions. They found it *exhausting*, and by the time they got home at the end of the day, a question like, "Would you rather we have peas or green beans with dinner?" often felt like the last straw. "I don't care about the stupid vegetables, just give me food and don't make me decide *any more things!*"

They were rescued by the following technique. When faced with a decision, they'd imagine "the Executive Director of the organization", and ask themselves, "What would 'the Executive Director of the organization' do?" Instead of making a decision, they'd make a *prediction* about the actions of that other person. Then, they'd just do whatever that person would do!

In my friend's case, they were trying to reduce decision fatigue. When I started trying it out myself, I was after a cure for something slightly different.

Imagine you're about to go bungee jumping off a high cliff. You know it's perfectly safe, and all you have to do is take a step forward, just like you've done every single time you've ever walked. But something is stopping you. The decision to step off the ledge is entirely yours, and you know you want to do it because this is why you're here. Yet here you are, still standing on the ledge.

You're scared. There's a battle happening in your brain. Part of you is going, "Just jump, it's easy, just do it!", while another part--the part in charge of your legs, apparently--is going, "NOPE. Nope nope nope nope NOPE." And you have this strange thought: "I wish someone would just push me so I don't have to decide."

Maybe you've been bungee jumping, and this is not at all how you responded to it. But I hope (for the sake of communication) that you've experienced this sensation in other contexts. Maybe when you wanted to tell someone that you loved them, but the phrase hovered just behind your lips, and you couldn't get it out. You almost wished it would tumble out of your mouth accidentally. "Just say it," you thought to yourself, and remained silent. For some reason, you were terrified of the decision, and inaction felt more like not deciding.

When I heard this story from my friend, I had social anxiety. I didn't have way more decisions than I knew how to handle, but I did find certain decisions terrifying, and was often paralyzed by them. For example, this always happened if someone I liked, respected, and wanted to interact with more asked to meet with them. It was pretty obvious to me that it was a good idea to say yes, but I'd agonize over the email endlessly instead of simply typing "yes" and hitting "send".

So here's what it looked like when I applied the technique. I'd be invited to a party. I'd feel paralyzing fear, and a sense of impending doom as I noticed that I likely believed going to the party was the right decision. Then, as soon as I felt that doom, I'd take a mental step backward and *not* try to force myself to decide. Instead, I'd imagine a version of myself *who wasn't scared*, and I'd *predict* what she'd do. If the party really *wasn't* a great idea, either because she didn't consider it worth my time or because she didn't actually anticipate me having any fun, she'd decide not to go. Otherwise, she'd decide to go. *I* would not decide. I'd just run my simulation of her, and see what she had to say. It was easy for her to think clearly about the decision, because she wasn't scared. And then I'd just defer to her.

Recently, I've noticed that there are all sorts of circumstances under which it helps to predict the decisions of a version of myself who doesn't have my current obstacle to rational decision making. Whenever I'm having a hard time thinking clearly about something because I'm angry, or tired, or scared, I can call upon imaginary Rational Brienne to see if she can do any better.

Example: I get depressed when I don't get enough sunlight. I was working inside where it was dark, and Eliezer noticed that I'd seemed depressed lately. So he told me he thought I should work outside instead. I was indeed a bit down and irritable, so my immediate response was to feel angry--that I'd been interrupted, that he was nagging me about getting sunlight *again*, and that I have this sunlight problem in the first place.

I started to argue with him, but then I stopped. I stopped because I'd noticed something. In addition to anger, I felt something like confusion. More complicated and specific than confusion, though. It's the feeling I get when I'm playing through familiar motions that have tended to lead to disutility. Like when you're watching a horror movie and the main character says, "Let's split up!" and you feel like, "Ugh, not this again. Listen, you're in a horror movie. If you split up, *you will die*. It happens every time." A familiar twinge of something being not quite right.

But even though I noticed the feeling, I couldn't get a handle on it. Recognizing that I really should make the decision to go outside instead of arguing--it was just too much for me. I was angry, and that severely impedes my introspective vision. And I knew that. I knew that familiar not-quite-right feeling meant something was preventing me from applying some of my rationality skills.

So, as I'd previously decided to do in situations like this, I called upon my simulation of non-angry Brienne.

She immediately got up and went outside.

To her, it was extremely obviously the right thing to do. So I just deferred to her (which I'd also previously decided to do in situations like this, and I knew it would only work in the future if I did it now too, ain't timeless decision theory great). I stopped arguing, got up, and went outside.

I was still *pissed*, mind you. I even felt myself rationalizing that I was doing it because going outside despite Eliezer being *wrong wrong wrong* is easier than arguing with him, and arguing with him isn't worth the effort. And then I told him as much over chat. (But not the "rationalizing" part; I wasn't fully conscious of that yet.)

But *I went outside*, right away, instead of wasting a bunch of time and effort first. My internal state was still in disarray, but I took the correct external actions.

This has happened a few times now. I'm still getting the hang of it, but it's working.

Imaginary Rational Brienne isn't magic. Her only available skills are the ones I have in fact picked up, so anything I've not learned, she can't implement. She still makes mistakes.

Her special strength is *constancy*.

In real life, all kinds of things limit my access to my own skills. In fact, the times when I most need a skill will very likely be the times when I find it hardest to access. For example, it's more important to consider the opposite when I'm really invested in believing something than when I'm not invested at all, but it's much harder to actually carry out the mental motion of "considering the opposite" when all the cognitive momentum is moving toward arguing single-mindedly for my favored belief.

The advantage of Rational Brienne (or, really, the Rational Briennes, because so far I've always ended up simulating a version of myself that's exactly the same except lacking whatever particular obstacle is relevant at the time) is that her access doesn't vary by situation. She can always use all of my tools all of the time.

I've been trying to figure out this constancy thing for quite a while. What do I do when I call upon my art as a rationalist, and just get a 404 Not Found? Turns out, "trying harder" doesn't do the trick. "No, really, I don't care that I'm scared, I'm going to think clearly about this. Here I go. I mean it this time." It seldom works.

I hope that it will one day. I would rather not have to rely on tricks like this. I hope I'll eventually just be able to go straight from noticing dissonance to re-orienting my whole mind so it's in line with the truth and with whatever I need to reach my goals. Or, you know, not experiencing the dissonance in the first place because I'm already doing everything right.

In the mean time, this trick seems pretty powerful.

Goal retention discussion with Eliezer

Although I feel that Nick Bostrom's new book "Superintelligence" is generally awesome and a well-needed milestone for the field, I do have one quibble: both he and Steve Omohundro appear to be more convinced than I am by the assumption that an AI will naturally tend to retain its goals as it reaches a deeper understanding of the world and of itself. I've written a short essay on this issue from my physics perspective, available at http://arxiv.org/pdf/1409.0813.pdf.

Eliezer Yudkowsky just sent the following extremely interesting comments, and told me he was OK with me sharing them here to spur a broader discussion of these issues, so here goes.

On Sep 3, 2014, at 17:21, Eliezer Yudkowsky <yudkowsky@gmail.com> wrote:

Hi Max! You're asking the right questions. Some of the answers we can give you, some we can't, few have been written up and even fewer in any well-organized way. Benja or Nate might be able to expound in more detail while I'm in my seclusion.

Very briefly, though:
The problem of utility functions turning out to be ill-defined in light of new discoveries of the universe is what Peter de Blanc named an "ontological crisis" (not necessarily a particularly good name, but it's what we've been using locally).

http://intelligence.org/files/OntologicalCrises.pdf

The way I would phrase this problem now is that an expected utility maximizer makes comparisons between quantities that have the type "expected utility conditional on an action", which means that the AI's utility function must be something that can assign utility-numbers to the AI's model of reality, and these numbers must have the further property that there is some computationally feasible approximation for calculating expected utilities relative to the AI's probabilistic beliefs. This is a constraint that rules out the vast majority of all completely chaotic and uninteresting utility functions, but does not rule out, say, "make lots of paperclips".

Models also have the property of being Bayes-updated using sensory information; for the sake of discussion let's also say that models are about universes that can generate sensory information, so that these models can be probabilistically falsified or confirmed. Then an "ontological crisis" occurs when the hypothesis that best fits sensory information corresponds to a model that the utility function doesn't run on, or doesn't detect any utility-having objects in. The example of "immortal souls" is a reasonable one. Suppose we had an AI that had a naturalistic version of a Solomonoff prior, a language for specifying universes that could have produced its sensory data. Suppose we tried to give it a utility function that would look through any given model, detect things corresponding to immortal souls, and value those things. Even if *the immortal-soul-detecting utility function works perfectly (it would in fact detect all immortal souls) this utility function will not detect anything in many (representations of) universes, and in particular it will not detect anything in the (representations of) universes we think have most of the probability mass for explaining our own world. In this case the AI's behavior is undefined until you tell me more things about the AI; an obvious possibility is that the AI would choose most of its actions based on low-probability scenarios in which hidden immortal souls existed that its actions could affect. (Note that even in this case the utility function is stable!)*

Since we don't know the final laws of physics and could easily be surprised by further discoveries in the laws of physics, it seems pretty clear that we shouldn't be specifying a utility function over exact physical states relative to the Standard Model, because if the Standard Model is even slightly wrong we get an ontological crisis. Of course there are all sorts of extremely good reasons we should not try to do this anyway, some of which are touched on in your draft; there just is no simple function of physics that gives us something good to maximize. See also Complexity of Value, Fragility of Value, indirect normativity, the whole reason for a drive behind CEV, and so on. We're almost certainly going to be using some sort of utility-learning algorithm, the learned utilities are going to bind to modeled final physics by way of modeled higher levels of representation which are known to be imperfect, and we're going to have to figure out how to preserve the model and learned utilities through shifts of representation. E.g., the AI discovers that humans are made of atoms rather than being ontologically fundamental humans, and furthermore the AI's multi-level representations of reality evolve to use a different sort of approximation for "humans", but that's okay because our utility-learning mechanism also says how to re-bind the learned information through an ontological shift.

This sorta thing ain't going to be easy which is the other big reason to start working on it well in advance. I point out however that this doesn't seem unthinkable in human terms. We discovered that brains are made of neurons but were nonetheless able to maintain an intuitive grasp on what it means for them to be happy, and we don't throw away all that info each time a new physical discovery is made. The kind of cognition we want does not seem inherently self-contradictory.

Three other quick remarks:

**) Natural selection is not a consequentialist, nor is it the sort of consequentialist that can sufficiently precisely predict the results of modifications that the basic argument should go through for its stability. The Omohundrian/Yudkowskian argument is not that we can take an arbitrary stupid young AI and it will be smart enough to self-modify in a way that preserves its values, but rather that most AIs that don't self-destruct will eventually end up at a stable fixed-point of coherent consequentialist values. This could easily involve a step where, e.g., an AI that started out with a neural-style delta-rule policy-reinforcement learning algorithm, or an AI that started out as a big soup of self-modifying heuristics, is "taken over" by whatever part of the AI first learns to do consequentialist reasoning about code. But this* *process doesn't repeat indefinitely; it stabilizes when there's a consequentialist self-modifier with a coherent utility function that can precisely predict the results of self-modifications. The part where this does happen to an initial AI that is under this threshold of stability is a big part of the problem of Friendly AI and it's why MIRI works on tiling agents and so on!*

**) Natural selection is not a consequentialist, nor is it the sort of consequentialist that can sufficiently precisely predict the results of modifications that the basic argument should go through for its stability. It built humans to be consequentialists that would value sex, not value inclusive genetic fitness, and not value being faithful to natural selection's optimization criterion. Well, that's dumb, and of course the result is that humans don't optimize for inclusive genetic fitness. Natural selection was just stupid like that. But that doesn't mean there's a generic process whereby an agent rejects its "purpose" in the light of exogenously appearing preference criteria. Natural selection's anthropomorphized "purpose" in making human brains is just not the same as the cognitive purposes represented in those brains. We're not talking about spontaneous rejection of internal cognitive purposes based on their causal origins failing to meet some exogenously-materializing criterion of validity. Our rejection of "maximize inclusive genetic fitness" is not an exogenous rejection of something that was explicitly represented in us, that we were explicitly being consequentialists for. It's a rejection of something that was never an explicitly represented terminal value in the first place. Similarly the stability argument for sufficiently advanced self-modifiers doesn't go through a step where the successor form of the AI reasons about the intentions of the previous step and respects them apart from its constructed utility function. So the lack of any universal preference of this sort is not a general obstacle to stable self-improvement.*

**) The case of natural selection does not illustrate a universal computational constraint, it illustrates something that we could anthropomorphize as a foolish design error. Consider humans building Deep Blue. We built Deep Blue to attach a sort of default value to queens and central control in its position evaluation function, but Deep Blue is still perfectly able to sacrifice queens and central control alike if the position reaches a checkmate thereby. In other words, although an agent needs crystallized instrumental goals, it is also perfectly reasonable to have an agent which never knowingly sacrifices the terminally defined utilities for the crystallized instrumental goals if the two conflict; indeed "instrumental value of X" is simply "probabilistic belief that X leads to terminal utility achievement", which is sensibly revised in the presence of any overriding information about the terminal utility. To put it another way, in a rational agent, the only way a loose generalization about instrumental expected-value can conflict with and trump terminal actual-value is if the agent doesn't know it, i.e., it does something that it reasonably expected to lead to terminal value, but it was wrong.*

This has been very off-the-cuff and I think I should hand this over to Nate or Benja if further replies are needed, if that's all right.

The Octopus, the Dolphin and Us: a Great Filter tale

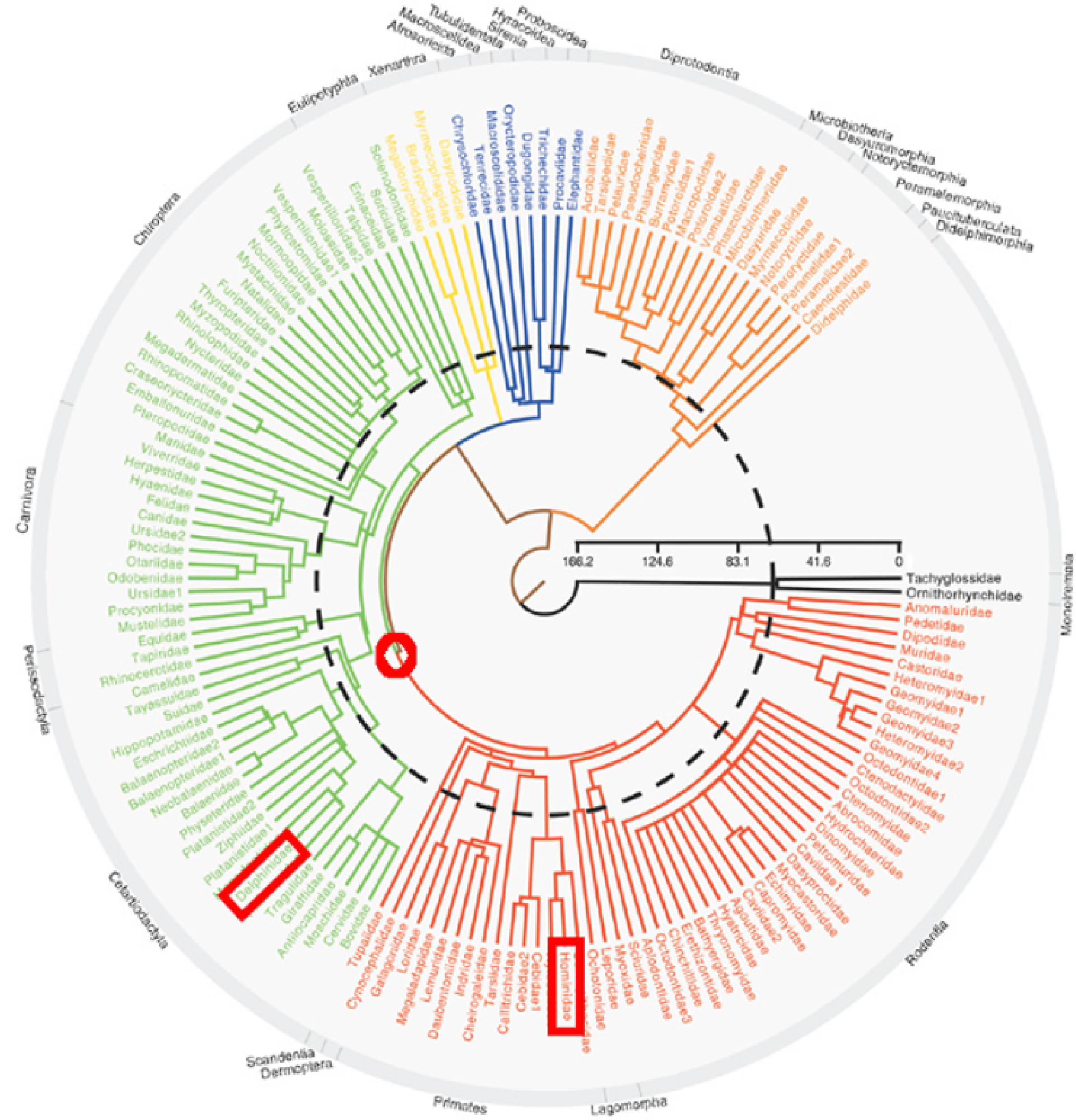
Is intelligence hard to evolve? Well, we're intelligent, so it must be easy... except that only an intelligent species would be able to ask that question, so we run straight into the problem of anthropics. Any being that asked that question would have to be intelligent, so this can't tell us anything about its difficulty (a similar mistake would be to ask "is most of the universe hospitable to life?", and then looking around and noting that everything seems pretty hospitable at first glance...).

Instead, one could point at the great apes, note their high intelligence, see that intelligence arises separately, and hence that it can't be too hard to evolve.

One could do that... but one would be wrong. The key test is not whether intelligence can arise *separately*, but whether it can arise *independently*. Chimpanzees, Bonobos and Gorillas and such are all "on our line": they are close to common ancestors of ours, which we would expect to be intelligent because we are intelligent. Intelligent species tend to have intelligent relatives. So they don't provide any extra information about the ease or difficulty of evolving intelligence.

To get independent intelligence, we need to go far from our line. Enter the smart and cute icon on many student posters: the dolphin.

Dolphins are certainly intelligent. And they are certainly far from our line. It seems hard to find a definite answer, but it seems that the last common ancestor of humans and dolphins was a small mammal existing during the reign of the dinosaurs. Humans and dolphins have been indicated by red rectangles, and their last common ancestor with a red circle.



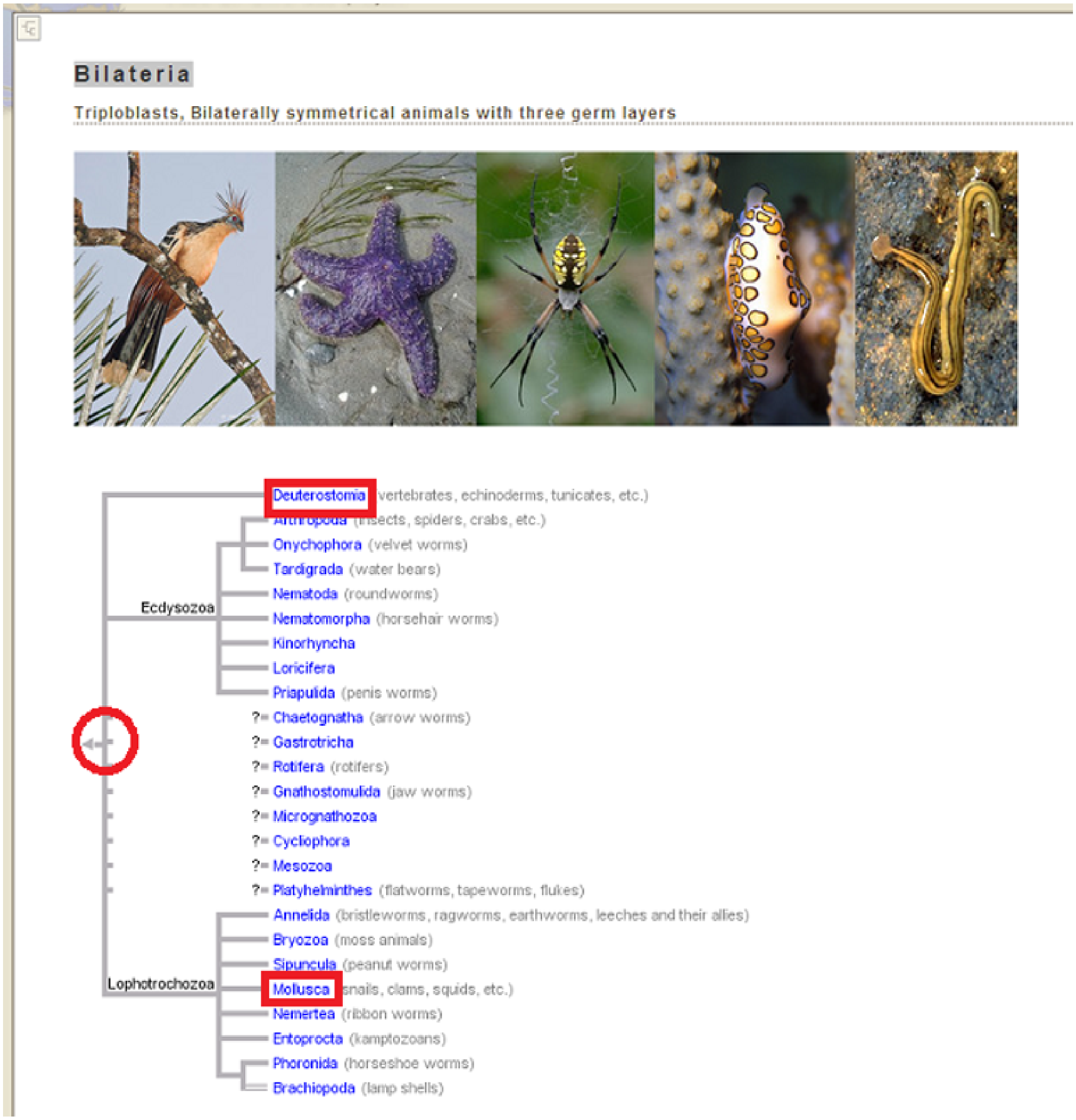
This red circle is well before the K-T boundary (indicated by the dotted line), hence represents a mammal living in the literal shadow of the dinosaurs.

We can apply a convergent evolution argument to this common ancestor. Thus, assuming that subsequent evolution was somewhat independent, getting from that common ancestor to dolphin level of intelligence is something that can happen relatively easily.

Can we go further? Well, what if we applied the argument twice? Let's bring in the most alien looking of the high-intelligence animals: the octopus.

Let's make the further assumption that our common ancestor with dolphins was dumber than the modern octopus. This doesn't seem a stretch seeing how intelligent the modern octopus can be, how minor in terms of ecological role the common dolphin-human ancestor must have been, and seeing the stupidity of many of the descendants of that common ancestor.

If we accept that assumption, we can then start looking for the common ancestor of humans and octopuses. Our two species are really far apart:



Bilateria

Triploblasts, Bilaterally symmetrical animals with three germ layers

Deuterostomia (vertebrates, echinoderms, tunicates, etc.)
Arthropoda (insects, spiders, crabs, etc.)
Onychophora (velvet worms)
Tardigrada (water bears)
Nematoda (roundworms)
Ecdysozoa
Nematomorpha (horsehair worms)
Kinorhyncha
Loricifera
Priapulida (penis worms)
?= Chaetognatha (arrow worms)
?= Gastrotricha
?= Rotifera (rotifers)
?= Gnathostomulida (jaw worms)
?= Micrognathozoa
?= Cyclophora
?= Mesozoa
?= Platyhelminthes (flatworms, tapeworms, flukes)
Annelida (bristleworms, ragworms, earthworms, leeches and their allies)
Bryozoa (moss animals)
Sipuncula (peanut worms)
Lophotrochozoa
Mollusca (snails, clams, squids, etc.)
Nemertea (ribbon worms)
Entoprocta (kamptozoans)
Phoronida (horseshoe worms)
Brachiopoda (lamp shells)

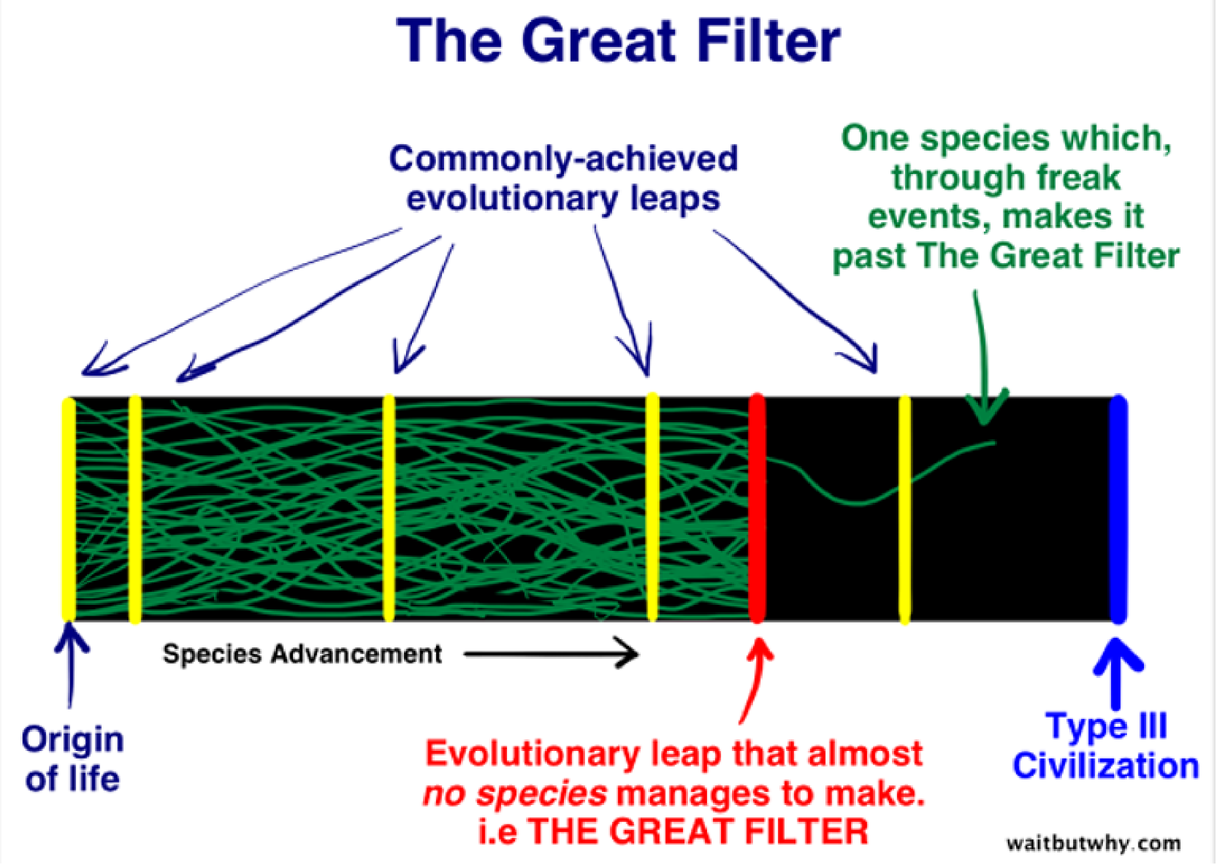
We therefore have to go back to around the last common ancestor of the Bilateria (creatures with bilateral symmetry, i.e. they have a front and a back end, as well as an upside and downside, and therefore a left and a right). This is the (speculative) urbilaterian. There are no known examples or fossils of it, which means that it was likely less than 1 cm in length. To quote Wikipedia: "The urbilaterian is often considered to have possessed a gut and internal organs, a segmented body and a centralised nervous system, as well as a biphasic life cycle (i.e. consisting of larvae and adults) and some features of embryonic development. However, this need not necessarily be the case." Very confusing, and with no information about intelligence level. However, since the organism was so small and since it was the ancestor of almost every animal alive today (including worms and Bryozoa), our best estimate would be that it's pretty stupid, with the simplest possible "brain".

Putting this all together, it seems evolutionarily easy to get from urbilatrian intelligence to Octopus intelligence, and from Octopus intelligence to dolphin intelligence - thus from urbilatrian to dolphin.

Note that this argument assumes that intelligence can be put on something like a linear scale. One could argue that Octopuses have low social intelligence, for instance. But then one could repeat the argument with distant animals with high social intelligence such as certain insects. Especially if one believe in a more general form of intelligence, it seems that this family of arguments could be used effectively to demonstrate dolphin-level intelligence emerging easily from very low levels of intelligence.

Application to the Great Filter

The Great Filter (related to the Fermi Paradox) is the argument that since we don't see any evidence of complex technological life in the universe, something must be preventing its emergence. At some point on the trajectory, something is culling almost all species.


The Great Filter
Commonly-achieved evolutionary leaps
One species which, through freak events, makes it past The Great Filter
Species Advancement
Origin of life
Evolutionary leap that almost *no species* manages to make. i.e THE GREAT FILTER
Type III Civilization
waitbutwhy.com

An "early" great filter wouldn't affect us: that means that we got through the filter already, it's in our past, so the emptiness among the stars doesn't say anything negative for us. A "late" great filter is bad news: that implies that few civilizations make it from technological civilization to star-spanning civilization, with bad results for us.

Note that AI is certainly not a great filter: an AI would likely expand through the universe itself

The real filter could be a combination of an early one and a late one, of course. But, unless the factors are exquisitely well-balanced, its likely that there is one location in civilizational development where most of the filter lies (ie where the probability of getting to the next stage is the lowest). Some possible locations for this could be:

- Life itself is unlikely (very early great filter).
- Life with a central nervous system is unlikely.
- **A lot of different possible locations for the great filter in between urbilatiran and dolphin intelligence.**
- Getting from dolphin to human intelligence is unlikely.
- Getting from human intelligence to technological civilization is unlikely (latest early filter).
- Getting from technological civilization to star-spanning civilization is unlikely.

These categories aren't of same size, of course - the first three are very diverse and large, for instance. Then what the evolutionary argument above says, is that the Great Filter in unlikely to be in the third, bolded category (which is in fact a multi-category).

For what it's worth, my personal judgement is that the filter lies before the creation of a central nervous system.

The Future of Humanity Institute could make use of your money

Many people have an incorrect view of the Future of Humanity Institute's funding situation, so this is a brief note to correct that; think of it as a spiritual successor to this post. As John Maxwell puts it, FHI is "one of the three organizations co-sponsoring LW [and] a group within the University of Oxford's philosophy department that tackles important, large-scale problems for humanity like how to go about reducing existential risk." (If you're not familiar with our work, this article is a nice, readable introduction, and our director, Nick Bostrom, wrote *Superintelligence*.) Though we are a research institute in an ancient and venerable institution, this does not guarantee funding or long-term stability.

Academic research is generally funded through grants, but because the FHI is researching important but unusual problems, and because this research is multi-disciplinary, we've found it difficult to attract funding from the usual grant bodies. This has meant that we’ve had to prioritise a certain number of projects that are not perfect for existential risk reduction, but that allow us to attract funding from interested institutions.

With more assets, we could both liberate our long-term researchers to do more "pure Xrisk" research, and hire or commission new experts when needed to look into particular issues (such as synthetic biology, the future of politics, and the likelihood of recovery after a civilization collapse).

We are not in any immediate funding crunch, nor are we arguing that the FHI would be a better donation target than MIRI, CSER, or the FLI. But any donations would be both gratefully received and put to effective use. If you'd like to, you can donate to FHI here. Thank you!

Newcomblike problems are the norm

This is crossposted from my blog. In this post, I discuss how Newcomblike situations are common among humans in the real world. The intended audience of my blog is wider than the readerbase of LW, so the tone might seem a bit off. Nevertheless, the points made here are likely new to many.

1

Last time we looked at Newcomblike problems, which cause trouble for Causal Decision Theory (CDT), the standard decision theory used in economics, statistics, narrow AI, and many other academic fields.

These Newcomblike problems may seem like strange edge case scenarios. In the Token Trade, a deterministic agent faces a perfect copy of themself, guaranteed to take the same action as they do. In Newcomb's original problem there is a perfect predictor Ω which knows exactly what the agent will do.

Both of these examples involve some form of "mind-reading" and assume that the agent can be perfectly copied or perfectly predicted. In a chaotic universe, these scenarios may seem unrealistic and even downright crazy. What does it matter that CDT fails when there are perfect mind-readers? There aren't perfect mind-readers. Why do we care?

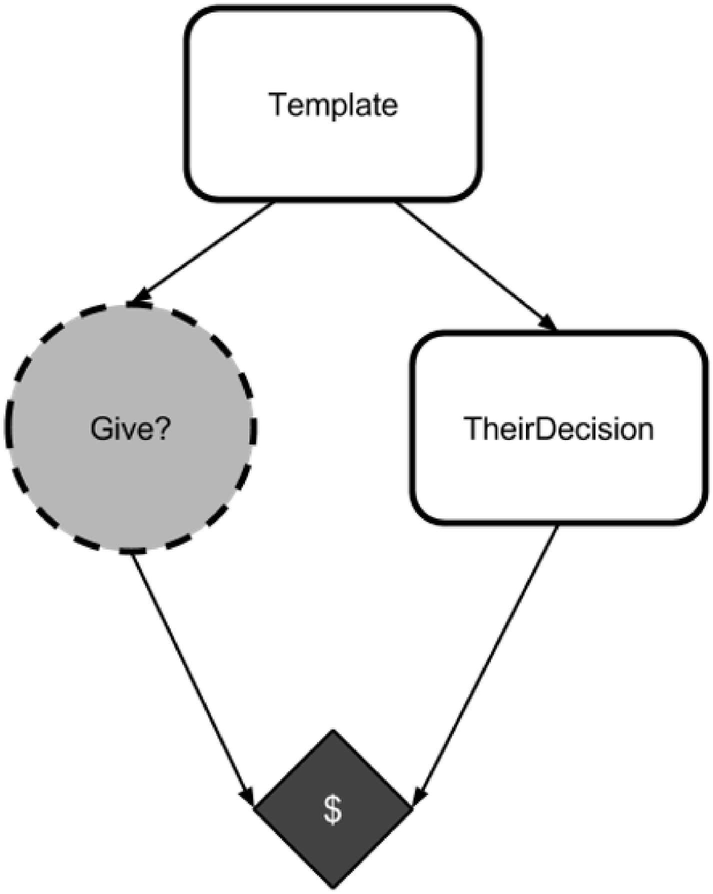
The reason that we care is this: *Newcomblike problems are the norm.* Most problems that humans face in real life are "Newcomblike".

These problems aren't limited to the domain of perfect mind-readers; rather, problems with perfect mind-readers are the domain where these problems are easiest to see. However, they arise naturally whenever an agent is in a situation where others have knowledge about its decision process via some mechanism that is not under its direct control.

2

Consider a CDT agent in a mirror token trade.

3



It knows that it and the opponent are generated from the same template, but it also knows that the opponent is causally distinct from it by the time it makes its choice. So it argues

> Either agents spawned from my template give their tokens away, or they keep their tokens. If agents spawned from my template give their tokens away, then I better keep mine so that I can take advantage of the opponent. If, instead, agents spawned from my template keep their tokens, then I had better keep mine, or otherwise I won't win any money at all.

It has failed, here, to notice that it can't choose separately from "agents spawned from my template" because it *is* spawned from its template. (That's not to say that it doesn't get to choose what to do. Rather, it has to be able to reason about the fact that whatever it chooses, so will its opponent choose.)

The reasoning flaw here is an inability to reason as if *past information* has given others *veridical knowledge* about what the agent *will* choose. This failure is particularly vivid in the mirror token trade, where the opponent is guaranteed to do *exactly* the same thing as the opponent. However, the failure occurs even if the veridical knowledge is partial or imperfect.

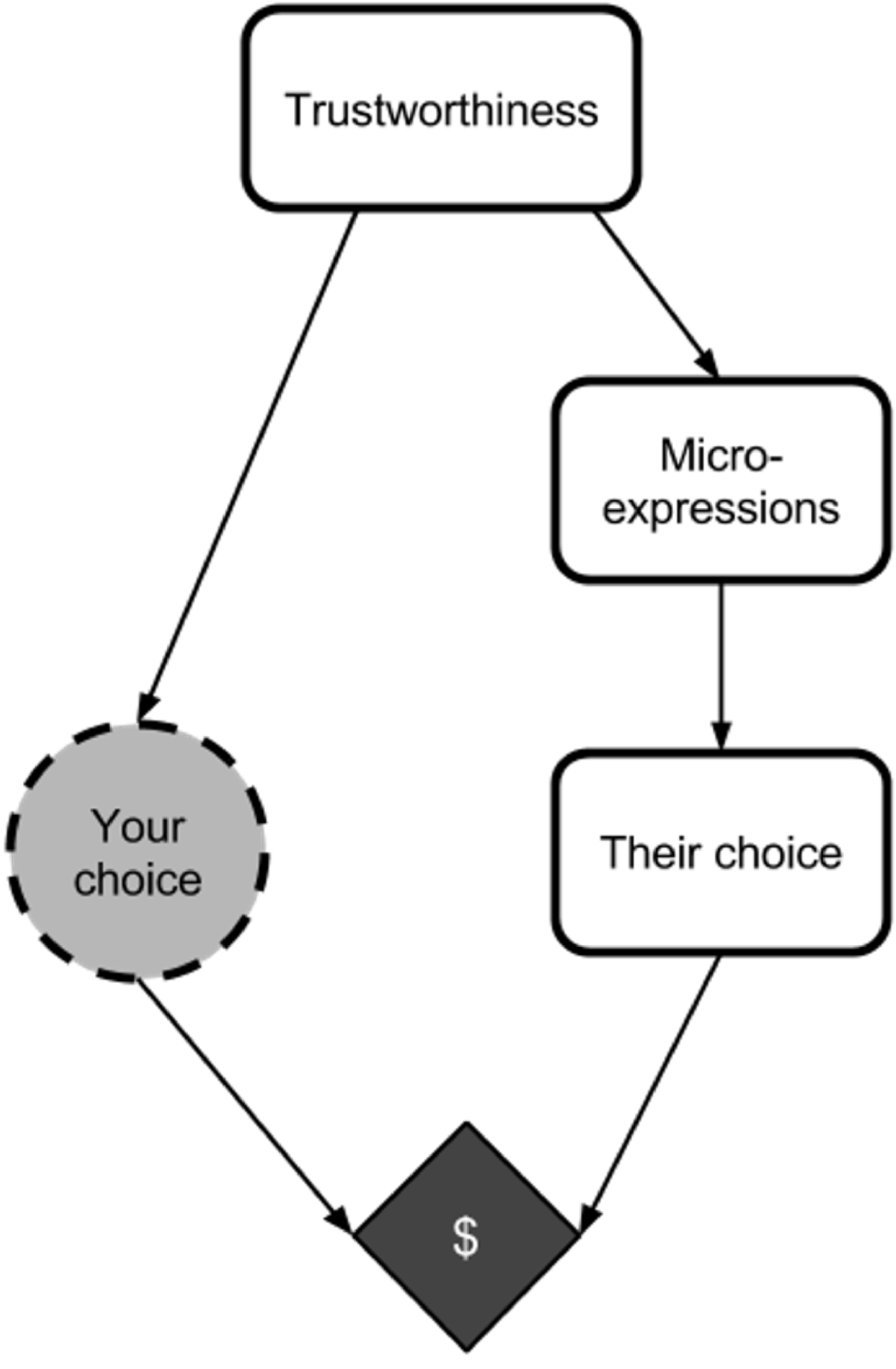
3

Humans trade partial, veridical, uncontrollable information about their decision procedures *all the time*.

Humans automatically make first impressions of other humans at first sight, almost instantaneously (sometimes before the person speaks, and possibly just from still images).

We read each other's microexpressions, which are generally uncontrollable sources of information about our emotions.

As humans, we have an impressive array of social machinery available to us that gives us gut-level, subconscious impressions of how trustworthy other people are.


Trustworthiness
Micro-expressions
Your choice
Their choice
$

Many social situations follow this pattern, and this pattern is a Newcomblike one.

All these tools can be fooled, of course. First impressions are often wrong. Con-men often seem trustworthy, and honest shy people can seem unworthy of trust. However, all of this social data is at least *correlated* with the truth, and that's all we need to give CDT trouble. Remember, CDT assumes that all nodes which are *causally* disconnected from it are *logically* disconnected from it: but if someone else gained information that correlates with how you *actually are* going to act in the future, then your interactions with them may be Newcomblike.

In fact, humans have a natural tendency to avoid "non-Newcomblike" scenarios. Human social structures use complex reputation systems. Humans seldom make big choices among themselves (who to hire, whether to become roommates, whether to make a business deal) before "getting to know each other". We automatically build complex social models detailing how we think our friends, family, and co-workers, make decisions.

When I worked at Google, I'd occasionally need to convince half a dozen team leads to sign off on a given project. In order to do this, I'd meet with each of them in person and pitch the project slightly differently, according to my model of what parts of the project most appealed to them. I was basing my actions off of how I expected them to make decisions: I was putting them in Newcomblike scenarios.

We constantly leak information about how we make decisions, and others constantly use this information. Human decision situations are Newcomblike *by default!* It's the *non*-Newcomblike problems that are simplifications and edge cases.

Newcomblike problems occur whenever knowledge about what decision you *will* make leaks into the environment. The knowledge doesn't have to be 100% accurate, it just has to be correlated with your eventual actual action (in such a way that if you were going to take a different action, then you would have leaked different information). When this information is available, and others use it to make their decisions, others put you into a Newcomblike scenario.

Information about what we're going to do is frequently leaking into the environment, via unconscious signaling and uncontrolled facial expressions or even just by habit — anyone following a simple routine is likely to act predictably.

4

Most real decisions that humans face are Newcomblike whenever other humans are involved. People are automatically reading unconscious or unintentional signals and using these to build models of how you make choices, and they're using those models to make *their* choices. These are precisely the sorts of scenarios that CDT cannot represent.

Of course, that's not to say that humans fail drastically on these problems. We don't: we repeatedly do well in these scenarios.

Some real life Newcomblike scenarios simply don't represent games where CDT has trouble: there are many situations where others in the environment have knowledge about how you make decisions, and are using that knowledge but in a way that does not affect your payoffs enough to matter.

Many more Newcomblike scenarios simply don't feel like decision problems: people present ideas to us in specific ways (depending upon their model of how we make choices) and most of us don't fret about how others would have presented us with different opportunities if we had acted in different ways.

And in Newcomblike scenarios that *do* feel like decision problems, humans use a wide array of other tools in order to succeed.

Roughly speaking, CDT fails when it gets stuck in the trap of "no matter what I signaled I should do [something mean]", which results in CDT sending off a "mean" signal and missing opportunities for higher payoffs. By contrast, humans tend to avoid this trap via other means: we place value on things like "niceness" for reputational reasons, we have intrinsic senses of "honor" and "fairness" which alter the payoffs of the game, and so on.

This machinery was not necessarily "designed" for Newcomblike situations. Reputation systems and senses of honor are commonly attributed to humans facing repeated scenarios (thanks to living in small tribes) in the ancestral environment, and it's possible to argue that CDT handles repeated Newcomblike situations well enough. (I disagree somewhat, but this is an argument for another day.)

Nevertheless, the machinery that allows us to handle repeated Newcomblike problems often seems to work in one-shot Newcomblike problems. Regardless of where the machinery came from, it still allows us to succeed in Newcomblike scenarios that we face in day-to-day life.

The fact that humans easily succeed, often via tools developed for repeated situations, doesn't change the fact that many of our day-to-day interactions have Newcomblike characteristics. Whenever an agent leaks information about their decision procedure on a communication channel that they do not control (facial microexpressions, posture, cadence of voice, etc.) that person is inviting others to put them in Newcomblike settings.

5

Most of the time, humans are pretty good at handling naturally arising Newcomblike problems. Sometimes, though, the fact that you're in a Newcomblike scenario *does* matter.

The games of Poker and Diplomacy are both centered around people controlling information channels that humans can't normally control. These games give particularly crisp examples of humans wrestling with situations where the environment contains leaked information about their decision-making procedure.

These are only games, yes, but I'm sure that any highly ranked Poker player will tell you that the lessons of Poker extend far beyond the game board. Similarly, I expect that highly ranked Diplomacy players will tell you that Diplomacy teaches you many lessons about how people broadcast the decisions that they're going to make, and that these lessons are invaluable in everyday life.

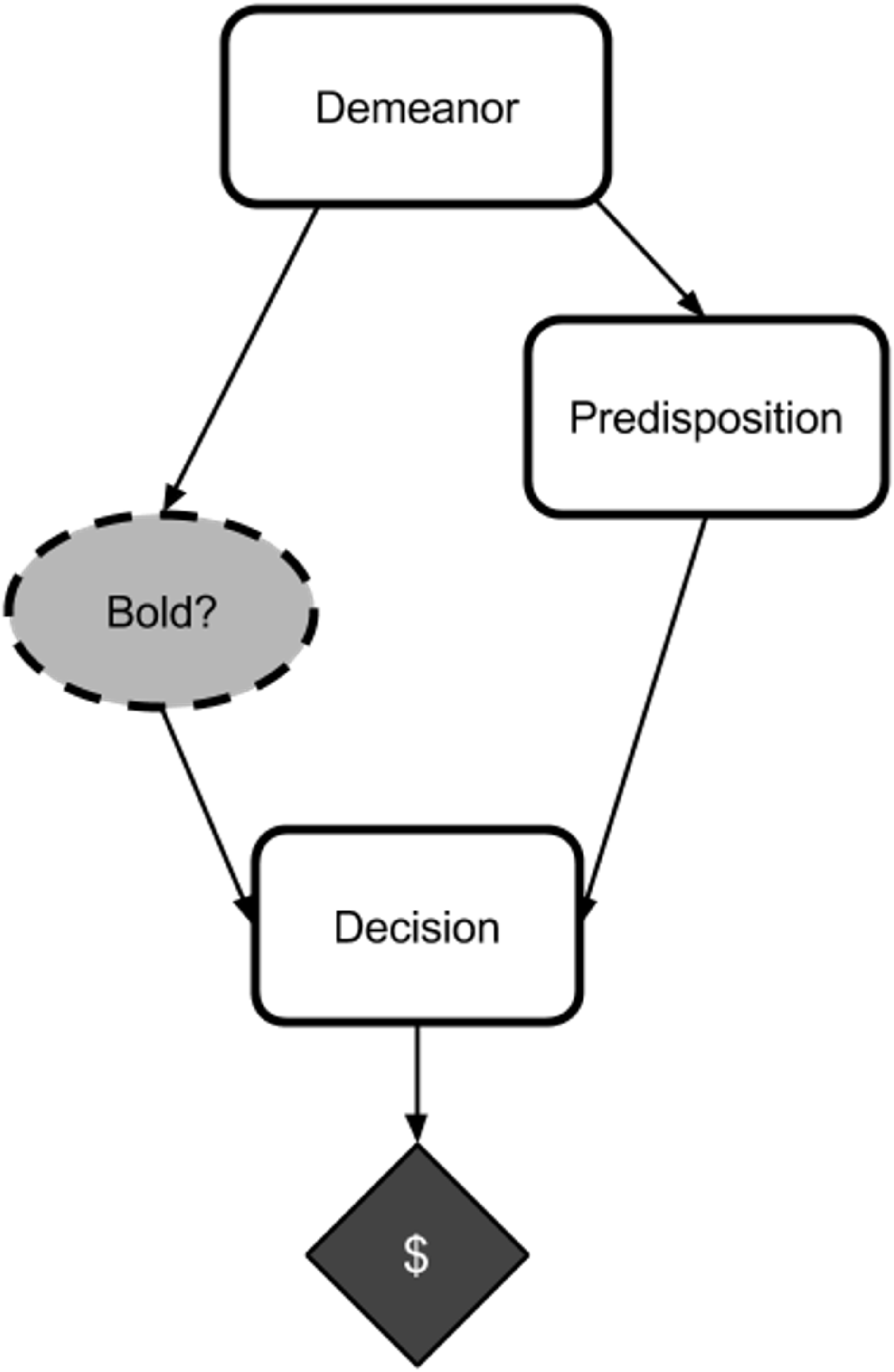
I am not a professional negotiator, but I further imagine that top-tier negotiators expend significant effort exploring how their mindsets are tied to their unconscious signals.

On a more personal scale, some very simple scenarios (like whether you can get let into a farmhouse on a rainy night after your car breaks down) are somewhat "Newcomblike".

I know at least two people who are unreliable and untrustworthy, and who blame the fact that they can't hold down jobs (and that nobody cuts them any slack) on bad luck rather than on their own demeanors. Both consistently believe that they are taking the best available action whenever they act unreliable and untrustworthy. Both brush off the idea of "becoming a sucker". Neither of them is capable of *acting* unreliable while *signaling* reliability. Both of them would benefit from *actually becoming trustworthy*.

Now, of course, people can't suddenly "become reliable", and akrasia is a formidable enemy to people stuck in these negative feedback loops. But nevertheless, you can see how this problem has a hint of Newcomblikeness to it.

In fact, recommendations of this form — "You can't signal trustworthiness unless you're trustworthy" — are common. As an extremely simple example, let's consider a shy candidate going in to a job interview. The candidate's demeanor (`confident` or `shy`) will determine the interviewer's predisposition `towards` or `against` the candidate. During the interview, the candidate may act either `bold` or `timid`. Then the interviewer decides whether or not to hire the candidate.


Demeanor
Predisposition
Bold?
Decision
$

If the candidate is confident, then they will get the job (worth $100,000) regardless of whether they are bold or timid. If they are shy and timid, then they will not get the job ($0). If, however, thy are shy and bold, then they will get laughed at, which is worth -$10. Finally, though, *a person who knows they are going to be timid will have a shy demeanor, whereas a person who knows they are going to be bold will have a confident demeanor*.

It may seem at first glance that it is better to be timid than to be bold, because timidness only affects the outcome if the interviewer is predisposed against the candidate, in which case it is better to be timid (and avoid being laughed at). However, if the candidate *knows* that they will reason like this (in the interview) then they will be shy *before* the interview, which will predispose the interviewer against them. By contrast, if the candidate precommits to being bold (in this simple setting) then the will get the job.

Someone reasoning using CDT might reason as follows when they're in the interview:

> I can't tell whether they like me or not, and I don't want to be laughed at, so I'll just act timid.

To people who reason like this, we suggest *avoiding causal reasoning* during the interview.

And, in fact, there are truckloads of self-help books dishing out similar advice. You can't reliably signal trustworthiness without *actually being* trustworthy. You can't reliably be charismatic without *actually caring* about people. You can't easily signal confidence without *becoming confident*. Someone who *cannot represent* these arguments may find that many of the benefits of trustworthiness, charisma, and confidence are unavailable to them.

Compare the advice above to our analysis of CDT in the mirror token trade, where we say "You can't keep your token while the opponent gives theirs away". CDT, which can't represent this argument, finds that the high payoff is unavailable to it. The analogy is exact: CDT fails to represent precisely this sort of reasoning, and yet this sort of reasoning is common and useful among humans.

6

That's not to say that CDT can't address these problems. A CDT agent that knows it's going to face the above interview would precommit to being bold — but this would involve using something *besides* causal counterfactual reasoning during the actual interview. And, in fact, this is precisely one of the arguments that I'm going to make in future posts: a sufficiently intelligent artificial system using CDT to reason about its choices would self-modify to stop using CDT to reason about its choices.

We've been talking about Newcomblike problems in a very human-centric setting for this post. Next post, we'll dive into the arguments about why an *artificial* agent (that doesn't share our vast suite of social signaling tools, and which lacks our shared humanity) may *also* expect to face Newcomblike problems and would therefore self-modify to stop using CDT.

This will lead us to more interesting questions, such as "what *would* it use?" (spoiler: we don't quite know yet) and "would it self-modify to fix all of CDT's flaws?" (spoiler: no).

What It's Like to Notice Things

Phenomenology

Phenomenology is the study of the structures of experience and consciousness. Literally, it is the study of "that which appears". The first time you look at a twig sticking up out of the water, you might be curious and ask, "What forces cause things to bend when placed in water?" If you're a curious phenomenologist, though, you'll ask things like, "Why does that twig in water appear as though bent? Do other things appear to bend when placed in water? Do all things placed in water appear to bend to the same degree? Are there things that do not appear to bend when placed in water? Does my perception of the bending depend on the angle or direction from which I observe the twig?"

Pehenomenology means breaking experience down to its more basic components, and being precise in our descriptions of what we actually observe, free of further speculation and assumption. A phenomenologist recognizes the difference between observing "a six-sided cube", and observing the three faces, at most, from which we extrapolate the rest.

I consider phenomenology to be a central skill of rationality. The most obvious example: You're unlikely to generate alternative hypotheses when the confirming observation and the favored hypothesis are one and the same in your experience of experience. The importance of phenomenology to rationality goes deeper than that, though. Phenomenology trains especially fine grained introspection. The more tiny and subtle are the thoughts you're aware of, the more precise can be the control you gain over the workings of your mind, and the faster can be your cognitive reflexes.

(I do not at all mean to say that you should go read Husserl and Heidegger. Despite their apparent potential for unprecedented clarity, the phenomenologists, without exception, seem to revel in obfuscation. It's probably not worth your time to wade through all of that nonsense. I've mostly read about phenomenology myself for this very reason.)

I've been doing some experimental phenomenology of late.

Noticing

I've noticed that rationality, in practice, depends on noticing. Some people have told me this is basically tautological, and therefore uninteresting. But if I'm right, I think it's likely very important to know, and to train deliberately.

The difference between seeing the twig as bent and seeing the twig as seeming bent may seem inane. It is not news that things that are bent tend to seem bent. Without that level of granularity in your observations, though, you may not notice that it could be possible for things to merely seem bent without being bent. When we're talking about something that may be ubiquitous to all applications of rationality, like noticing, it's worth taking a closer look at the contents of our experiences.

Many people talk about "noticing confusion", because Eliezer's written about it. Really, though, every successful application of a rationality skill begins with noticing. In particular, applied rationality is founded on noticing opportunities and obstacles. (To be clear, I'm making this up right this moment, so as far as I know it's not a generally agreed-upon thing. That goes for nearly everything in this post. I still think it's true.) You can be the most technically skilled batter in the world, and it won't help a bit if you consistently fail to notice when the ball whizzes by you--if you miss the opportunities to swing. And you're not going to run very many bases if you launch the ball straight at an opposing catcher--if you're oblivious to the obstacles.

It doesn't matter how many techniques you've learned if you miss all the opportunities to apply them, and fail to notice the obstacles when they get in your way. Opportunities and obstacles are everywhere. We can only be as strong as our ability to notice the ones that will make a difference.

Inspired by Whales' self-experiment in noticing confusion, I've been practicing noticing things. Not difficult or complicated things, like noticing confusion, or noticing biases. I've just been trying to get a handle on noticing, full stop. And it's been interesting.

Noticing Noticing

What does it mean to notice something, and what does it feel like?

I started by checking to see what I expected it to feel like to notice that it's raining, just going from memory. I tried for a split-second prediction, to find what my brain automatically stored under "noticing rain". When I thought about noticing rain, I got this sort of vague impression of rainyness, which included few sensory details and was more of an overall rainy feeling. My brain tried to tell me that "noticing rain" meant "being directly acquainted with rainyness", in much the same way that it tries to tell me it's experiencing a cube when it's actually only experiencing a pattern of light and shadows I interpret as three faces.

Then, I waited for rain. It didn't take long, because I'm in North Carolina for the month. (This didn't happen last time I was in North Carolina, so perhaps I just happened to choose The One Valley of Eternal Rain.)

The real "noticing rain" turned out to be a response to the physical sensations concurrent with the first raindrop falling on my skin. I did eventually have an "abstract rainyness feeling", but that happened a full two seconds later. My actual experience went like this.

It was cloudy and humid. This was not at the forefront of my attention, but it slowly moved in that direction as the temperature dropped. I was fairly focused on reading a book.

(I'm a little baffled by the apparent gradient between "not at all conscious of x" and "fully aware of x". I don't know how that works, but I experience the difference between being a little aware of the sky being cloudy and being focused on the patterns of light in the clouds, as analogous to the difference between being very-slightly-but-not-uncomfortably warm and burning my hand on the stove.)

My awareness of something like an "abstract rainyness feeling" moved further toward consciousness as the wind picked up. Suddenly--and the suddenness was an important part of the experience--I felt something like a cool, dull pin-prick on my arm. I looked at it, saw the water, and recognized it as a raindrop. Over the course of about half a second, several sensations leapt forward into full awareness: the darkness of my surroundings, the humidity in the air, the dark grey-blueness of the sky, the sound of rain on leaves like television static, the scent of ozone and damp earth, the feeling of cool humid wind on my face, and the word "rain" in my internal monologue.

I think it is that sudden leaping forward of many associated sensations that I would call "noticing rain".

After that, I felt a sort of mental step backward--though it was more like a zooming out or sliding away than a discrete step--from the sensations, and then a feeling of viewing them from the outside. There was a sensation of the potential to access other memories of times when it's rained.

(Sensations of potential are fascinating to me. I noticed a few weeks ago that after memorizing a list of names and faces, I could predict in the first half second of seeing the face whether or not I'd be able to retrieve the name in the next five seconds. Before I actually retrieved the name. What??? I don't know either.)

Only then did all of it resolve into the more distant and abstract "feeling of rainyness" that I'd predicted before. The resolution took four times as long as the simultaneous-leaping-into-consciousness-of-related-sensations that I now prefer to call "noticing", and ten times as long as the first-raindrop-pin-prick, which I think I'll call the "noticing trigger" if it turns out to be a general class of pre-noticing experiences.

("Can you really distinguish between 200 and 500 milliseconds?" Yes, but it's an acquired skill. I spent a block of a few minutes every day for a month, then several blocks a day for about a week, doing this Psychomotor Vigiliance Task when I was gathering data for the polyphasic sleep experiment. (No, I'm sorry, to the best of my knowledge Leverage has not yet published anything on the results of this. Long story short: Everyone who wasn't already polyphasic is still not polyphasic today.) It gives you fast feedback on simple response time. I'm not sure if it's useful for anything else, but it comes in handy when taking notes on experiences that pass very quickly.)

Noticing Environmental Cues

My second experiment was in repeated noticing. This is more closely related to rationality as habit cultivation.

Can I get better at noticing something just by practicing?

I was trying to zoom in on the experience of noticing itself, so I wanted something as simple as possible. Nothing subtle, nothing psychological, and certainly nothing I might be motivated to ignore. I wanted a straightforward element of my physical environment. I'm out in the country and driving around for errands and such about once a day, so I went with "red barn roofs".

I had an intuition that I should give myself some outward sign of having noticed, lest I not notice that I noticed, and decided to snap my fingers every time I noticed a red barn roof.

On the first drive, I noticed one red barn roof. That happened when I was almost at my destination and I thought, "Oh right, I'm supposed to be noticing red barn roofs, oops" then started actively searching for them.

Noticing a red barn roof while searching for it feels very different from noticing rain while reading a book. With the rain, it felt sort of like waking up, or like catching my name in an overheard conversation. There was a complete shift in what my brain was doing. With the barn roof, it was like I had a box with a red-barn-roof-shaped hole, and it felt like completion when a I grabbed a roof and dropped it through the hole. I was prepared for the roof, and it was a smaller change in the contents of consciousness.

I noticed two on the way back, also while actively searching for them, before I started thinking about something else and became oblivious.

I thought that maybe there weren't enough red barn roofs, and decided to try noticing red roofs of all sorts of buildings the next day. This, it turns out, was the correct move.

On day two of red-roof-noticing, I got lots of practice. I noticed around fifteen roofs on the way to the store, and around seven on the way back. By the end, I was not searching for the roofs as intently as I had been the day before, but I was still explicitly thinking about the project. I was still aware of directing my eyes to spend extra time at the right level in my field of vision to pick up roofs. It was like waving the box around and waiting for something to fall in, while thinking about how to build boxes.

I went out briefly again on day two, and on the way back, I noticed a red roof while thinking about something else entirely. Specifically, I was thinking about the possibility of moving to Uruguay, and whether I knew enough Spanish to survive. In the middle of one of those unrelated thoughts, my eyes moved over a barn roof and stayed there briefly while I had the leaping-into-consciousness experience with respect to the sensations of redness, recognizing something as shaped like a building, and feeling the impulse to snap my fingers. It was like I'd been wearing the box as a hat to free up my hands, and I'd forgotten about it. And then, with a heavy ker-thunk, the roof became my new center of attention.

And oh my gosh, it was so exciting! It sounds so absurd in retrospect to have been excited about noticing a roof. But I was! It meant I'd successfully installed a new cognitive habit to run in the background. On purpose. "Woo hoo! Yeah!" (I literally said that.)

On the third day, I noticed TOO MANY red roofs. I followed the same path to the store as before, but I noticed somewhere between twenty and thirty red roofs. I got about the same number going back, so I think I was catching nearly all the opportunities to notice red roofs. (I'd have to do it for a few days to be sure.) There was a pattern to noticing, where I'd notice-in-the-background, while thinking about something else, the first roof, and then I'd be more specifically on the lookout for a minute or two after that, before my mind wandered back to something other than roofs. I got faster over time at returning to my previous thoughts after snapping my fingers, but there were still enough noticed roofs to intrude uncomfortably upon my thoughts. It was getting annoying.

So I decided to switch back to only noticing the red roofs of barns in particular.

Extinction of the more general habit didn't take very long. It was over by the end of my next fifteen minute drive. For the first three times I saw a roof, I rose my hand a little to snap my fingers before reminding myself that I don't care about non-barns anymore. The next couple times I didn't raise my hand, but still forcefully reminded myself of my disinterest in my non-barns. The promotion of red roofs into consciousness got weaker with each roof, until the difference between seeing a non-red non-barn roof and a red non-barn roof was barely perceptible. That was my drive to town today.

On the drive back, I noticed about ten red barn roofs. Three I noticed while thinking about how to install habits, four while thinking about the differences between designing exercises for in-person workshops and designing exercises to put in books, and three soon enough after the previous barn to probably count as "searching for barns".

So yes, for at least some things, it seems I can get better at noticing them my by practicing.

What These Silly Little Experiments Are Really About

My plan is to try noticing an internal psychological phenomenon next, but still something straightforward that I wouldn't be motivated not to notice. I probably need to try a couple things to find something that works well. I might go with "thinking the word 'tomorrow' in my internal monologue", for example, or possibly "wondering what my boyfriend is thinking about". I'll probably go with something more like the first, because it is clearer, and zooms in on "noticing things inside my head" without the extra noise of "noticing things that are relatively temporally indiscrete", but the second is actually a useful thing to notice.

Most of the useful things to notice are a lot less obvious than "thinking the word 'tomorrow' in my internal monologue". From what I've learned so far, I think that for "wondering what my boyfriend is thinking about", I'll need to pick out a couple of very specific, instantaneous sensations that happen when I'm curious what my boyfriend is thinking about. I expect that to be a repetition of the rain experiment, where I predict what it will feel like, then wait 'til I can gather data in real time. Once I have a specific trigger, I can repeat the red roof experiment to catch the tiny moments when I wonder what he's thinking. I might need to start with a broader category, like "notice when I'm thinking about my boyfriend", get used to noticing those sensations, and then reduce the set of sensations I'm watching out for to things that happen only when I'm curious what my boyfriend is thinking.

After that, I imagine I'll want to practice with different kinds of actions I can take when I notice a trigger. (If you've never heard of Implementation Intentions, I suggest trying them out.) So far, I've used the physical action of snapping my fingers. That was originally for clarity in recognizing the noticing, but it's also a behavioral response to a trigger. I could respond with a psychological behavior instead of a physical one, like "imagining a carrot". A useful response to noticing that I'm curious about what my boyfriend is thinking would be "check to see if he's busy" and then "say, 'What are you thinking about?'"

See, this "noticing" thing sounds boringly simple at first, and not worth much consideration in the art of rationality. Even in his original "noticing confusion" post, Eliezer really talked more about recognizing the implications of confusion than about the noticing itself.

Noticing is more complicated than it seems at first, and it's easy to mix it up with responding. There's a whole sub-art to noticing, and I really think that deliberate practice is making me better at it. Responses can be hard. It's essential to make noticing as effortless as possible. Then you can break the noticing and the responding apart, and you can recognize reality even before you know what to do with it.

Bayesianism for humans: "probable enough"

(followup to What Bayesianism has taught me? and Bayesianism for Humans)

There are two insights from Bayesianism which occurred to me and which I hadn't seen anywhere else before.
I like lists in the two posts linked above, so for the sake of completeness, I'm going to add my two cents to a public domain. Second penny is here.

"Probable enough"

> *When you have eliminated the impossible, whatever remains is often more improbable than your having made a mistake in one of your impossibility proofs.*
> Steven Kaas

Bayesian way of thinking introduced me to the idea of "hypothesis which is probably isn't true, but probable enough to rise to the level of conscious attention" — in other words, to the situation when P(H) is notable but less than 50%.

Looking back, I think that the notion of taking seriously something which you don't think is true was alien to me. Hence, everything was either probably true or probably false; things from the former category were over-confidently certain, and things from the latter category were barely worth thinking about.

This model was correct, but only in a formal sense.

Suppose you are living in Gotham, the city famous because of it's crime rate and it's masked (and well-funded) vigilante, Batman. Recently you had read *The Better Angels of Our Nature: Why Violence Has Declined* by Steven Pinker, and according to some theories described here, Batman isn't good for Gotham at all.

Now you know, for example, the theory of Donald Black that "*crime is, from the point of view of the perpetrator, the pursuit of justice*". You know about idea that in order for crime rate to drop, people should perceive their law system as legitimate. You suspect that criminals beaten by Bats don't perceive the act as a fair and regular punishment for something bad, or an attempt to defend them from injustice; instead the act is perceived as a round of bad luck. So, the criminals are busy plotting their revenge, not internalizing civil norms.

You believe that if you send your copy of book (with key passages highlighted) to the person connected to Batman, Batman will change his ways and Gotham will become much more nice in terms of homicide rate.

So you are trying to find out Batman's secret identity, and there are 17 possible suspects. Derek Powers looks like a good candidate: he is wealthy, and has a long history of secretly delegating illegal-violence-including tasks to his henchmen; however, his motivation is far from obvious. You estimate P(Derek Powers employs Batman) as 20%. You have very little information about other candidates, like Ferris Boyle, Bruce Wayne, Roland Daggett, Lucius Fox or Matches Malone, so you assign an equal 5% to everyone else.

In this case you should pick Derek Powers as your best guess when forced to name only one candidate (for example, if you forced to send the book to someone today), but also you should be aware that your guess is 80% likely to be wrong. When making expected utility calculations, you should take Derek Powers more seriously than Lucius Fox, but only by 15% more seriously.

In other words, you should take *maximum a posteriori probability* hypothesis into account while not deluding yourself into thinking that now you understand everything or nothing at all. Derek Powers hypothesis probably isn't *true;* but it is *useful*.

Sometimes I find it easier to reframe question from "what hypothesis is true?" to "what hypothesis is probable enough?". Now it's totally okay that your pet theory isn't probable but still probable enough, so doubt becomes easier. Also, you are aware that your pet theory is likely to be wrong (and this is nothing to be sad about), so the alternatives come to mind more naturally.

These "probable enough" hypothesis can serve as a very concise summaries of state of your knowledge when you simultaneously outline the general sort of evidence you've observed, and stress that you aren't really sure. I like to think about it like a rough, qualitative and more System1-friendly variant of Likelihood ratio sharing.

Planning Fallacy

The original explanation of planning fallacy (proposed by Kahneman and Tversky) is about people focusing on a most optimistic scenario when asked about typical one (instead of trying to do an Outside VIew). If you keep the distinction between "probable" and "probable enough" in mind, you can see this claim in a new light.

Because the most optimistic scenario *is* the most probable *and* the most typical one, in a certain sense.

The illustration, with numbers pulled out of thin air, goes like this: so, you want to visit a museum.

The first thing you need to do is to get dressed and take your keys and stuff. Usually (with 80% probability) you do this very quick, but there is a weak possibility of your museum ticket having been devoured by an entropy monster living on your computer table.

The second thing is to catch bus. Usually (p = 80%), bus is on schedule, but sometimes it can be too early or too late. After this, the bus could (20%) or could not (80%) get stuck in a traffic jam.

Finally, you need to find a museum building. You've been there before once, so you sorta remember your route, yet still could be lost with 20% probability.

And there you have it: P(everything is fine) = 40%, and probability of every other scenario is 10% or even less. "Everything is fine" is *probable enough,* yet likely to be false. Supposedly, humans pick MAP hypothesis and then forget about every other scenario in order to save computations.

Also, "everything is fine" is a good description of your plan. If your friend asks you, "so how are you planning to get to the museum?", and you answer "well, I catch the bus, get stuck in a traffic jam for 30 agonizing minutes, and then just walk from here", your friend is going to get a completely wrong idea about dangers of your journey. So, in a certain sense, "everything is fine" is a typical scenario.

Maybe it isn't human inability to pick the most likely scenario which should be blamed. Maybe it is false assumption that "most likely == likely to be correct" which contributes to this ubiquitous error.

In this case you would be better off having picked the "something will go wrong, and I will be late", instead of "everything will be fine".

So, sometimes you are interested in the best specimen out of your hypothesis space, sometimes you are interested in a most likely thingy (and it doesn't matter how vague it would be), and sometimes there are no shortcuts, and you have to do an actual expected utility calculation.

Unpopular ideas attract poor advocates: Be charitable

Unfamiliar or unpopular ideas will tend to reach you via proponents who:

- ...hold extreme interpretations of these ideas.
- ...have unpleasant social characteristics.
- ...generally come across as cranks.

The basic idea: It's unpleasant to promote ideas that result in social sanction, and frustrating when your ideas are met with indifference. Both situations are more likely when talking to an ideological out-group. Given a range of positions on an in-group belief, who will decide to promote the belief to outsiders? On average, it will be those who believe the benefits of the idea are large relative to in-group opinion (extremists), those who view the social costs as small (disagreeable people), and those who are dispositionally drawn to promoting weird ideas (cranks).

I don't want to push this pattern too far. This isn't a refutation of any particular idea. There are reasonable people in the world, and some of them even express their opinions in public, (in spite of being reasonable). And sometimes the truth will be unavoidably unfamiliar and unpopular, etc. But there are also...

Some benefits that stem from recognizing these selection effects:

- It's easier to be charitable to controversial ideas, when you recognize that you're interacting with people who are terribly suited to persuade you. I'm not sure "steelmanning" is the best idea (trying to present the best argument for an opponent's position). Based on the extremity effect, another technique is to construct a much diluted version of the belief, and then try to steelman the diluted belief.
- If your group holds fringe or unpopular ideas, you can avoid these patterns when you want to influence outsiders.
- If you want to learn about an afflicted issue, you might ignore the public representatives and speak to the non-evangelical instead (you'll probably have to start the conversation).
- You can resist certain polarizing situations, in which the most visible camps hold extreme and opposing views. This situation worsens when those with non-extreme views judge the risk of participation as excessive, and leave the debate to the extremists (who are willing to take substantial risks for their beliefs). This leads to the perception that the current camps represent the only valid positions, which creates a polarizing loop. Because this is a sort of coordination failure among non-extremists, knowing to covertly look for other non-vocal moderates is a first step toward a solution. (Note: Sometimes there really aren't any moderates.)
- Related to the previous point: You can avoid exaggerating the ideological unity of a group based on the group's leadership, or believing that the entire group has some obnoxious trait present in the leadership. (Note: In things like elections and war, the views of the leadership are what you care about. But you still don't want to be confused about other group members.)

I think the first benefit listed is the most useful.

To sum up: An unpopular idea will tend to get poor representation for social reasons, which will makes it seem like a worse idea than it really is, even granting that many unpopular ideas are unpopular for good reason. So when you encounter a idea that seem unpopular, you're probably hearing about it from a sub-optimal source, and you should try to be charitable towards the idea before dismissing it.

Talking to yourself: A useful thinking tool that seems understudied and underdiscussed

I have returned from a particularly fruitful Google search, with unexpected results.

My question was simple. I was pretty sure that talking to myself aloud makes me temporarily better at solving problems that need a lot of working memory. It is a thinking tool that I find to be of great value, and that I imagine would be of interest to anyone who'd like to optimize their problem solving. I just wanted to collect some evidence on that, make sure I'm not deluding myself, and possibly learn how to enhance the effect.

This might be just lousy Googling on my part, but the evidence is surprisingly unclear and disorganized. There are at least three seperate Wiki pages for it. They don't link to each other. Instead they present the distinct models of three seperate fields: autocommunication in communication studies, semiotics and other cultural studies, intrapersonal communication ("self-talk" redirects here) in anthropology and (older) psychology and private speech in developmental psychology. The first is useless for my purpose, the second mentions "may increase concentration and retention" with no source, the third confirms my suspicion that this behavior boosts memory, motivation and creativity, but it only talks about children.

Google Scholar yields lots of sports-related results for "self-talk" because it can apparently improve the performance of athletes and if there's something that obviously needs the optimization power of psychology departments, it is competitive sports. For "intrapersonal communication" it has papers indicating it helps in language acquisition and in dealing with social anxiety. Both are dwarfed by the results for "private speech", which again focus on children. There's very little on "autocommunication" and what is there has nothing to do with the functioning of individual minds.

So there's a bunch of converging pieces of evidence supporting the usefulness of this behavior, but they're from several seperate fields that don't seem to have noticed each other very much. How often do you find that?

Let me quickly list a few ways that I find it plausible to imagine talking to yourself could enhance rational thought.

- It taps the phonological loop, a distinct part of working memory that might otherwise sit idle in non-auditory tasks. More memory is always better, right?
- Auditory information is retained more easily, so making thoughts auditory helps remember them later.
- It lets you commit to thoughts, and build upon them, in a way that is more powerful (and slower) than unspoken thought while less powerful (but quicker) than action. (I don't have a good online source for this one, but *Inside Jokes* should convince you, and has lots of new cognitive science to boot.)
- System 1 does seem to understand language, especially if it does not use complex grammar - so this might be a useful way for results of System 2 reasoning to be propagated. Compare affirmations. Anecdotally, whenever I'm

starting a complex task, I find stating my intent out loud makes a *huge* difference in how well the various submodules of my mind cooperate.
- It lets separate parts of your mind communicate in a fairly natural fashion, slows each of them down to the speed of your tongue and makes them not interrupt each other so much. (This is being used as a psychotherapy method.) In effect, your mouth becomes a kind of talking stick in their discussion.

All told, if you're talking to yourself you should be more able to solve complex problems than somebody of your IQ who doesn't, although somebody of your IQ with a pen and a piece of paper should still outthink both of you.

Given all that, I'm surprised this doesn't appear to have been discussed on LessWrong. Honesty: Beyond Internal Truth comes close but goes past it. Again, this might be me failing to use a search engine, but I think this is worth more of our attention that it has gotten so far.

I'm now almost certain talking to myself is useful, and I already find hindsight bias trying to convince me I've always been so sure. But I wasn't - I was suspicious because talking to yourself is an early warning sign of schizophrenia, and is frequent in dementia. But in those cases, it might simply be an autoregulatory response to failing working memory, not a pathogenetic element. After all, its memory enhancing effect is what the developmental psychologists say the kids use it for. I do expect social stigma, which is why I avoid talking to myself when around uninvolved or unsympathetic people, but my solving of complex problems tends to happen away from those anyway so that hasn't been an issue really.

So, what do you think? Useful?