Best of LessWrong: April 2015

1. 16 types of useful predictions
2. Even better cryonics – because who needs nanites anyway?
3. Sapiens
4. Concept Safety: Producing similar AI-human concept spaces
5. How to sign up for Alcor cryo
6. Replacing guilt
7. Failing with abandon
8. The stamp collector

Best of LessWrong: April 2015

1. 16 types of useful predictions
2. Even better cryonics – because who needs nanites anyway?
3. Sapiens
4. Concept Safety: Producing similar AI-human concept spaces
5. How to sign up for Alcor cryo
6. Replacing guilt
7. Failing with abandon
8. The stamp collector

16 types of useful predictions

How often do you make predictions (either about future events, or about information that you don't yet have)? If you're a regular Less Wrong reader you're probably familiar with the idea that you should make your beliefs pay rent by saying, "Here's what I expect to see if my belief is correct, and here's how confident I am," and that you should then update your beliefs accordingly, depending on how your predictions turn out.

And yet… my impression is that few of us actually make predictions on a regular basis. Certainly, for me, there has always been a gap between how useful I think predictions are, in theory, and how often I make them.

I don't think this is just laziness. I think it's simply not a trivial task to find predictions to make that will help you improve your models of a domain you care about.

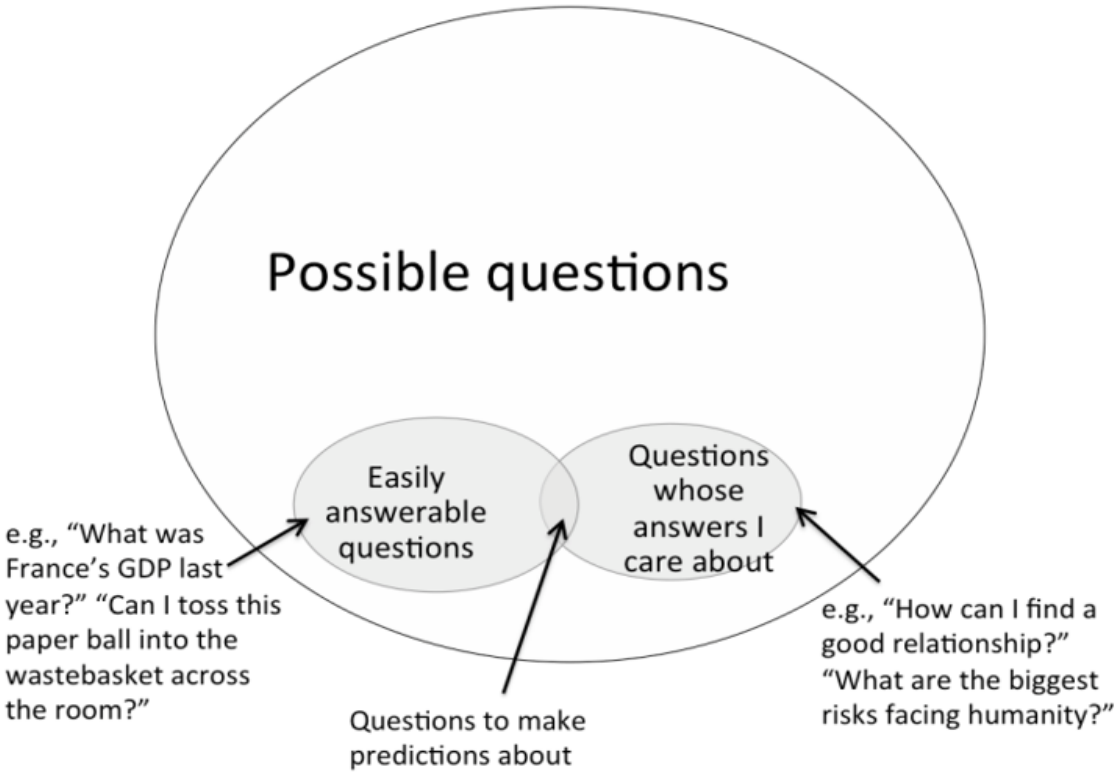
At this point I should clarify that there are two main goals predictions can help with:

1. **Improved Calibration** (e.g., realizing that I'm only correct about Domain X 70% of the time, not 90% of the time as I had mistakenly thought).
2. **Improved Accuracy** (e.g., going from being correct in Domain X 70% of the time to being correct 90% of the time)

If your goal is just to become better calibrated in general, it doesn't much matter what kinds of predictions you make. So calibration exercises typically grab questions with easily obtainable answers, like "How tall is Mount Everest?" or "Will Don Draper die before the end of Mad Men?" See, for example, the Credence Game, Prediction Book, and this recent post. And calibration training really does work.

But even though making predictions about trivia will improve my general calibration skill, it won't help me improve my models of the world. That is, it won't help me become more *accurate*, at least not in any domains I care about. If I answer a lot of questions about the heights of mountains, I might become more accurate about that topic, but that's not very helpful to me.

So I think the difficulty in prediction-making is this: The set {questions whose answers you can easily look up, or otherwise obtain} is a small subset of all possible questions. And the set {questions whose answers I care about} is also a small subset of all possible questions. And the intersection between those two subsets is much smaller still, and not easily identifiable. As a result, prediction-making tends to seem too effortful, or not fruitful enough to justify the effort it requires.



But the intersection's not empty. It just requires some strategic thought to determine which answerable questions have some bearing on issues you care about, or -- approaching the problem from the opposite direction -- how to take issues you care about and turn them into answerable questions.

I've been making a concerted effort to hunt for members of that intersection. Here are 16 types of predictions that I personally use to improve my judgment on issues I care about. (I'm sure there are plenty more, though, and hope you'll share your own as well.)

1. **Predict how long a task will take you.** This one's a given, considering how common and impactful the planning fallacy is.
   Examples: "How long will it take to write this blog post?" "How long until our company's profitable?"
2. **Predict how you'll feel in an upcoming situation.** Affective forecasting – our ability to predict how we'll feel – has some well known flaws.
   Examples: "How much will I enjoy this party?" "Will I feel better if I leave the house?" "If I don't get this job, will I still feel bad about it two weeks later?"
3. **Predict your performance on a task or goal.**
   One thing this helps me notice is when I've been trying the same kind of approach repeatedly without success. Even just the act of making the prediction can spark the realization that I need a better game plan.
   Examples: "Will I stick to my workout plan for at least a month?" "How well will this event I'm organizing go?" "How much work will I get done today?" "Can I successfully convince Bob of my opinion on this issue?"
4. **Predict how your audience will react to a particular social media post (on Facebook, Twitter, Tumblr, a blog, etc.)**.

This is a good way to hone your judgment about how to create successful content, as well as your understanding of your friends' (or readers') personalities and worldviews.
Examples: "Will this video get an unusually high number of likes?" "Will linking to this article spark a fight in the comments?"

5. **When you try a new activity or technique, predict how much value you'll get out of it.**
I've noticed I tend to be inaccurate in both directions in this domain. There are certain kinds of life hacks I feel sure are going to solve all my problems (and they rarely do). Conversely, I am overly skeptical of activities that are outside my comfort zone, and often end up pleasantly surprised once I try them.
Examples: "How much will Pomodoros boost my productivity?" "How much will I enjoy swing dancing?"
6. **When you make a purchase, predict how much value you'll get out of it.**
Research on money and happiness shows two main things: (1) as a general rule, money doesn't buy happiness, but also that (2) there are a bunch of exceptions to this rule. So there seems to be lots of potential to improve your prediction skill here, and spend your money more effectively than the average person.
Examples: "How much will I wear these new shoes?" "How often will I use my club membership?" "In two months, will I think it was worth it to have repainted the kitchen?" "In two months, will I feel that I'm still getting pleasure from my new car?"
7. **Predict how someone will answer a question about themselves**.
I often notice assumptions I'm been making about other people, and I like to check those assumptions when I can. Ideally I get interesting feedback both about the object-level question, and about my overall model of the person.
Examples: "Does it bother you when our meetings run over the scheduled time?" "Did you consider yourself popular in high school?" "Do you think it's okay to lie in order to protect someone's feelings?"
8. **Predict how much progress you can make on a problem in five minutes.**
I often have the impression that a problem is intractable, or that I've already worked on it and have considered all of the obvious solutions. But then when I decide (or when someone prompts me) to sit down and brainstorm for five minutes, I am surprised to come away with a promising new approach to the problem.
Example: "I feel like I've tried everything to fix my sleep, and nothing works. If I sit down now and spend five minutes thinking, will I be able to generate at least one new idea that's promising enough to try?"
9. **Predict whether the data in your memory supports your impression.**
Memory is awfully fallible, and I have been surprised at how often I am unable to generate specific examples to support a confident impression of mine (or how often the specific examples I generate actually contradict my impression).
Examples: "I have the impression that people who leave academia tend to be glad they did. If I try to list a bunch of the people I know who left academia, and how happy they are, what will the approximate ratio of happy/unhappy people be?"
"It feels like Bob never takes my advice. If I sit down and try to think of examples of Bob taking my advice, how many will I be able to come up with?"
10. **Pick one expert source and predict how they will answer a question.**
This is a quick shortcut to testing a claim or settling a dispute.
Examples: "Will Cochrane Medical support the claim that Vitamin D promotes hair growth?" "Will Bob, who has run several companies like ours, agree that our starting salary is too low?"

11. **When you meet someone new, take note of your first impressions of him. Predict how likely it is that, once you've gotten to know him better, you will consider your first impressions of him to have been accurate.**
    A variant of this one, suggested to me by CFAR alum Lauren Lee, is to make predictions about someone before you meet him, based on what you know about him ahead of time.
    Examples: "All I know about this guy I'm about to meet is that he's a banker; I'm moderately confident that he'll seem cocky." "Based on the one conversation I've had with Lisa, she seems really insightful – I predict that I'll still have that impression of her once I know her better."
12. **Predict how your Facebook friends will respond to a poll**.
    Examples: I often post social etiquette questions on Facebook. For example, I recently did a poll asking, "If a conversation is going awkwardly, does it make things better or worse for the other person to comment on the awkwardness?" I confidently predicted most people would say "worse," and I was wrong.
13. **Predict how well you understand someone's position by trying to paraphrase it back to him**.
    The illusion of transparency is pernicious.
    Examples: "You said you think running a workshop next month is a bad idea; I'm guessing you think that's because we don't have enough time to advertise, is that correct?"
    "I know you think eating meat is morally unproblematic; is that because you think that animals don't suffer?"
14. **When you have a disagreement with someone, predict how likely it is that a neutral third party will side with you after the issue is explained to her**.
    For best results, don't reveal which of you is on which side when you're explaining the issue to your arbiter.
    Example: "So, at work today, Bob and I disagreed about whether it's appropriate for interns to attend hiring meetings; what do you think?"
15. **Predict whether a surprising piece of news will turn out to be true**.
    This is a good way to hone your bullshit detector and improve your overall "common sense" models of the world.
    Examples: "This headline says some scientists uploaded a worm's brain -- after I read the article, will the headline seem like an accurate representation of what really happened?"
    "This viral video purports to show strangers being prompted to kiss; will it turn out to have been staged?"
16. **Predict whether a quick online search will turn up any credible sources supporting a particular claim.**
    Example: "Bob says that watches always stop working shortly after he puts them on – if I spend a few minutes searching online, will I be able to find any credible sources saying that this is a real phenomenon?"

I have one additional, general thought on how to get the most out of predictions:

Rationalists tend to focus on the importance of objective metrics. And as you may have noticed, a lot of the examples I listed above fail that criterion. For example, "Predict whether a fight will break out in the comments? Well, there's no objective way to say whether something officially counts as a 'fight' or not…" Or, "Predict whether I'll be able to find credible sources supporting X? Well, who's to say what a credible source is, and what counts as 'supporting' X?"

And indeed, objective metrics are preferable, all else equal. But all else isn't equal. Subjective metrics are much easier to generate, and they're far from useless. Most of the time it will be clear enough, once you see the results, whether your prediction basically came true or not -- even if you haven't pinned down a precise, objectively measurable success criterion ahead of time. Usually the result will be a common sense "yes," or a common sense "no." And sometimes it'll be "um...sort of?", but that can be an interestingly surprising result too, if you had strongly predicted the results would point clearly one way or the other.

Along similar lines, I usually don't assign numerical probabilities to my predictions. I just take note of where my confidence falls on a qualitative "very confident," "pretty confident," "weakly confident" scale (which might correspond to something like 90%/75%/60% probabilities, if I had to put numbers on it).

There's probably some additional value you can extract by writing down quantitative confidence levels, and by devising objective metrics that are impossible to game, rather than just relying on your subjective impressions. But in most cases I don't think that additional value is worth the cost you incur from turning predictions into an onerous task. In other words, don't let the perfect be the enemy of the good. Or in other other words: the biggest problem with your predictions right now is that they don't exist.

Even better cryonics – because who needs nanites anyway?

***Abstract**: in this post I propose a protocol for cryonic preservation (with the central idea of using high pressure to prevent water from expanding rather than highly toxic cryoprotectants), which I think has a chance of being non-destructive enough for us to be able to preserve and then resuscitate an organism with modern technologies. In addition, I propose a simplified experimental protocol for a shrimp (or other small model organism (building a large pressure chamber is hard) capable of surviving in very deep and cold waters; shrimp is a nice trade-off between the depth of habitat and the ease of obtaining them on market), which is simple enough to be doable in a small lab or well-equipped garage setting.*

Are there obvious problems with this, and how can they be addressed?

Is there a chance to pitch this experiment to a proper academic institution, or garage it is?

Originally posted here.

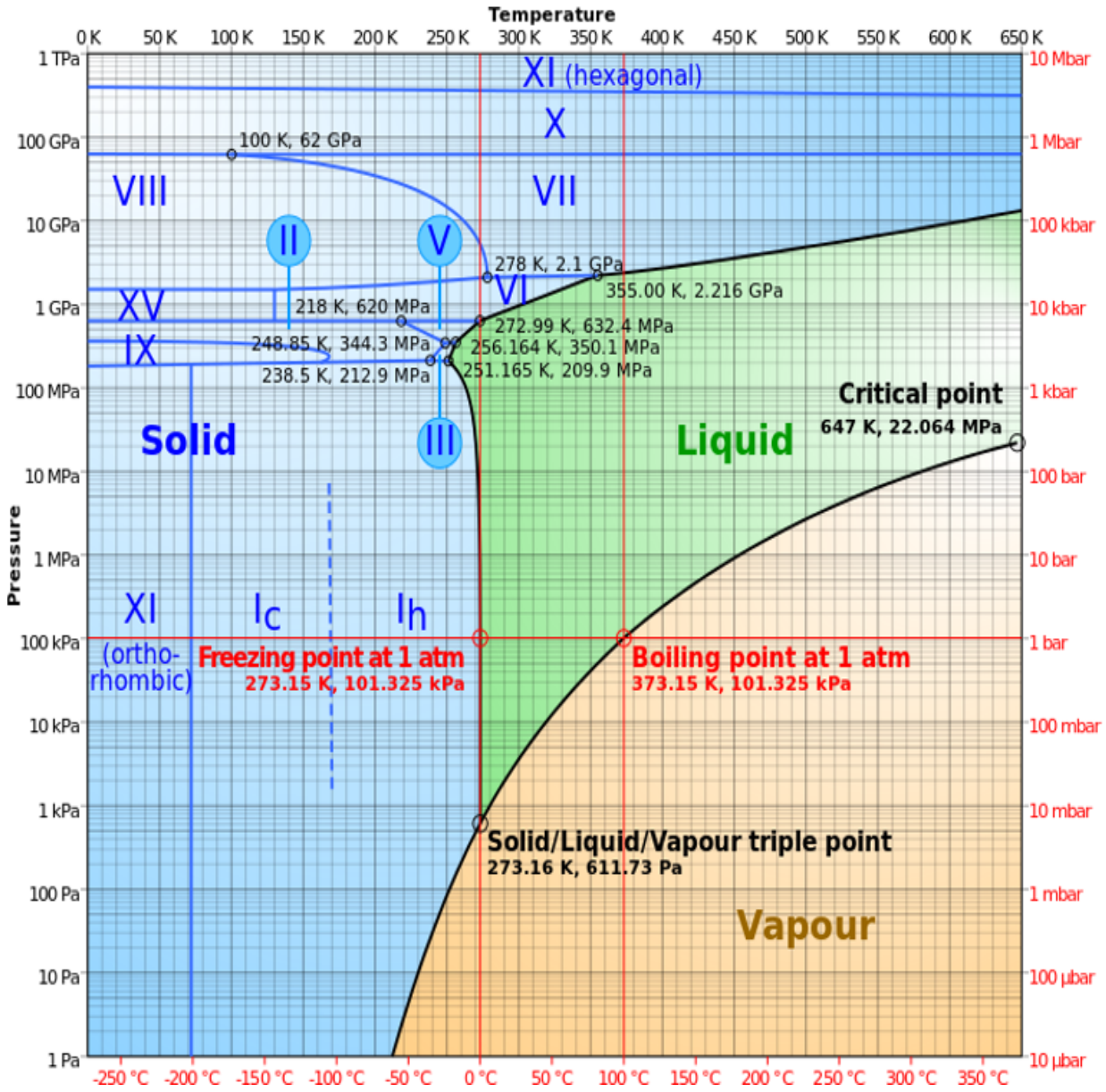
I do think that the odds of ever developing advanced nanomachines and/or brain scanning on molecular level plus algorithms for reversing information distortion - everything you need to undo the damage from conventional cryonic preservation and even to some extent that of brain death, according to its modern definition, if wasn't too late when the brain was preserved - for currently existing cryonics to be a bet worth taking. This is dead serious, and it's an actionable item.

Less of an action item: what if the future generations actually build quantum Bayesian superintelligence, close enough in its capabilities to Solomonoff induction, at which point even a mummified brain or the one preserved in formalin would be enough evidence to restore its original state? Or what if they invent read-only time travel, and make backups of everyone's mind right before they died (at which point it becomes indistinguishable from the belief in afterlife existing right now)? Even without time travel, they can just use a Universe-sized supercomputer to simulate every singe human physically possible, and naturally of of them is gonna be you. But aside from the obvious identity issues (and screw the timeless identity), that relies on unknown unknowns with uncomputable probabilities, and I'd like to have as few leaps of faith and quantum suicides in my life as possible.

So although vitrification right after diagnosed brain death relies on far smaller assumptions, and if totally worth doing - let me reiterate that: go sign up for cryonics - it'd be much better if we had preservation protocols so non-destructive that we could actually freeze a living human, and then bring them back alive. If nothing else, that would hugely increase the public outreach, grant the patient (rather than cadaver) status to the preserved, along with the human rights, get it recognized as a medical procedure covered by insurance or single payer, allow doctors to initiate the preservation of a dying patient before the brain death (again: I think everything short of information-theoretic death should potentially be reversible, but why take chances?), allow suffering patient opt for preservation rather than euthanasia (actually, I think it should be done right now: why on earth would anyone allow a person to do something that's guaranteed to kill them, but not allowed to do something that maybe will kill, or

maybe will give the cure?), or even allow patients suffering from degrading brain conditions (e.g. Alzheimer's) to opt for preservation before their memory and personality are permanently destroyed.

Let's fix cryonics! First of all, why can't we do it on living organisms? Because of heparin poisoning - every cryoprotectant efficient enough to prevent the formation of ice crystals is a strong enough poison to kill the organism (leave alone that we can't even saturate the whole body with it - current technologies only allow to do it for the brain alone). But without cryoprotectants the water will expand upon freezing, and break the cells. But there's another way to prevent this. Under pressure above 350 MPa [water slightly shrinks upon freezing](#) rather than expanding:



So that's basically that: **the key idea is to freeze (and keep) everything under pressure**. Now, there are some tricks to that too.

It's not easy to put basically any animal, especially a mammal, under 350 MPa (which is 3.5x higher than in Mariana Trench). At this point even Trimix becomes toxic. Basically the only remaining solution is total liquid ventilation, which has one problem: it has never been applied successfully to a human. There's one fix to that I see: as far as I can tell, no one has ever attempted to do perform it under high pressure, and the attempts were basically failing because of the insufficient solubility of oxygen and carbon dioxide in perfluorocarbons. Well then, let's increase the pressure! Namely, go to 3 MPa on Trimix, which is doable, and only then switch to TLV, whose efficiency is improved by the higher gas solubility under high pressure. But there's another solution too. If you just connect a cardiopulmonary bypass (10 hours should be enough for the whole procedure), you don't need the surrounding liquid to even be breathable - it can just be saline. CPB also solves the problem of surviving the period after the cardiac arrest (which will occur at around 30 centigrade) but before the freezing happens - you can just keep the blood circulating and delivering oxygen.

Speaking of hypoxia, even with the CPB it's still a problem. You positively don't want the blood to circulate when freezing starts, lest it act like an abrasive water cutter. It's not that much of a problem under near-freezing temperatures, but still. Fortunately, this effect can be mitigated by administering insulin first (yay, it's the first proper academic citation in this post! Also yay, I thought about this before I even discovered that it's actually true). This makes sense: if oxygen is primarily used to metabolize glucose, less glucose means less oxygen consumed, and less damage done by hypoxia. Then there's another thing: on the phase diagram you can see that before going into the area of high temperature ice at 632 MPa, freezing temperature actually dips down to roughly -30 centigrade at 209~350 MPa. That would allow to really shut down metabolism for good when water is still liquid, and blood can be pumped by the CPB. From this point we have two ways. First, we can do the normal thing, and start freezing very slowly, so minimize the formation of ice crystals (even though they're smaller than the original water volume, they may still be sharp). Second, we can increase the pressure. That would lead to near-instantaneous freezing everywhere, thus completely eliminating the problem of hypoxia - before the freezing, blood still circulated, and freezing is very quick - way faster than can ever be achieved even by throwing a body into liquid helium under normal pressure. Video evidence suggests that quick freezing of water leads to the formation of a huge number of crystals, which is bad, but I don't know near-instantaneous freezing from supercooled state and near-instantaneous freezing upon raising the pressure will lead to the same effect. More experiments are needed, preferably not on humans.

So here is my preservation protocol:

1. Anesthetize a probably terminally ill, but still conscious person.
2. Connect them to a cardiopulmonary bypass.
3. Replacing their blood with perfluorohexane is not necessary, since we seem to be already doing a decent job at having medium-term (several days) cardiopulmonary bypasses, but that could still help.
4. Submerge them in perfluorohexane, making sure that no air bubbles are left.
5. Slowly raise the ambient pressure to 350 MPa (~3.5kBar) without stopping the bypass.
6. Apply a huge dose of insulin to reduce all their metabolic processes.
7. Slowly cool them to -30 centigrade (at which point, given such pressure, water is still liquid), while increasing the dose of insulin, and raising the oxygen supply to the barely subtoxic level.
8. Slowly raise the pressure to 1 GPa (~10kBar), at which point the water solidifies, but does so with shrinking rather than expanding. Don't cutoff the blood

circulation until the moment when ice crystals starts forming in the blood/perfluorohexane flow.
9. Slowly lower the temperature to -173 centigrade or lower, as you wish.

And then back:

1. Raise the temperature to -20 centigrade.
2. Slowly lower the pressure to 350 MPa, at which point ice melts.
3. Start artificial blood circulation with a barely subtoxic oxygen level.
4. Slowly raise the temperature to +4 centigrade.
5. Slowly lower the pressure to 1 Bar.
6. Drain the ambient perfluorohexane and replace it with pure oxygen. Attach and start a medical ventilator.
7. Slowly raise the temperature to +32 centigrade.
8. Apply a huge dose of epinephrine and sugar, while transfusing the actual blood (preferably autotransfusion), to restart the heart.
9. Rejoice.

I claim that this protocol allows you freeze a living human to an arbitrarily low temperature, and then bring them back alive without brain damage, thus being the first true victory over death.

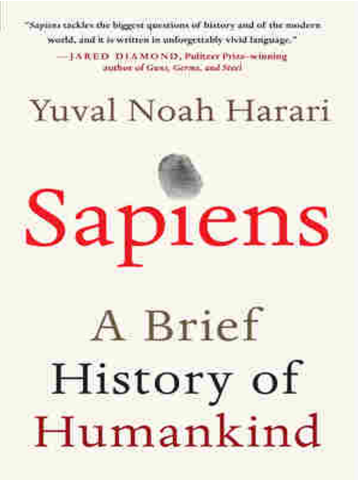
But let's start with something easy and small, like a shrimp. They already live in water, so there's no need to figure out the protocol for putting them into liquid. And they're already adapted to live under high pressure (no swim bladders or other cavities). And they're already adapted to live in cold water, so they should be expected to survive further cooling.

Small ones can be about 1 inch big, so let's be safe and use a 5cm-wide cylinder. To form ice III we need about 350MPa, which gives us 350e6 * 3.14 * 0.025^2 / 9.8 = 70 tons or roughly 690kN of force. Applying it directly or with a lever is unreasonable, since 70 tons of bending force is a lot even for steel, given the 5cm target. Block and tackle system is probably a good solution - actually, two of them, on each side of a beam used for compression, so we have 345 kN per system. And it looks like you can buy 40~50 ton manual hoists from alibaba, though I have no idea about their quality.

I'm not sure to which extent Pascal's law applies to solids, but if it does, the whole setup can be vastly optimized by creating a bottle neck for the pistol. One problem is that we can no longer assume that water in completely incompressible - it had to be compressed to about 87% its original volume - but aside from that, 350MPa per a millimeter thick rod is just 28kg. To compress a 0.05m by 0.1m cylinder to 87% its original volume we need to pump extra 1e-4 m^3 of water there, which amounts to 148 meters of movement, which isn't terribly good. 1cm thick rod, on the other hand, would require almost 3 tons of force, but will move only 1.5 meters. Or the problem of applying the constant pressure can be solved by enclosing the water in a plastic bag, and filling the rest of chamber with a liquid with a lower freezing point, but the same density. Thus, it is guaranteed that all the time it takes the water to freeze, it is under uniform external pressure, and then it just had nowhere to go.

Alternatively, one can just buy a 90'000 psi pump and 100'000 psi tubes and vessels, but let's face it: it they don't even list the price on their website, you probably don't even wanna know it. And since no institutions that can afford this thing seem to be interested in cryonics research, we'll have to stick to makeshift solutions (until at least the shrimp thing works, which would probably give in a publication in Nature, and enough academic recognition for proper research to start).

Sapiens



This is a section-by-section summary and review of Sapiens: A Brief History of Humankind by Yuval Noah Harari. It's come up on Less Wrong before in the context of Death is Optional, a conversation the author had with Daniel Kahneman about the book, and seems like an accessible introduction to many of the concepts underlying the LW perspective on history and the future. Anyone who's thought about Moloch will find many of the same issues discussed here, and so I'll scatter links to Yvain throughout. I'll discuss several of the points that I thought were interesting and novel, or at least had a novel perspective and good presentation.

A history as expansive as this one necessarily involves operating on higher levels of abstraction. The first section expresses this concisely enough to quote in full:

> About 13.5 billion years ago, matter, energy, time and space came into being in what is known as the Big Bang. The story of these fundamental features of our universe is called physics.
>
> About 300,000 years after their appearance, matter and energy started to coalesce into complex structures, called atoms, which then combined into molecules. The story of atoms, molecules and their interactions is called chemistry.
>
> About 3.8. billion years ago, on a planet called Earth, certain molecules combined to form particularly large and intricate structures called organisms. The story of organisms is called biology.
>
> About 70,000 years ago, organisms belonging to the species Homo sapiens started to form even more elaborate structures called cultures. The subsequent development of these human cultures is called history.
>
> Three important revolutions shaped the course of history: the Cognitive Revolution kick-started history about 70,000 years ago. The Agricultural Revolution sped it up about 12,000 years ago. The Scientific Revolution, which got under way only 500 years ago, may well end history and start something completely different. This book tells the story of how these three revolutions have affected humans and their fellow organisms.

The main charm of the book, as I see it, is what I would call the "view from Mars," or what others might call the "outside view" or "view from nowhere." We live in the middle of the Scientific Revolution, of history, biology, chemistry, and physics, and it is hard to imagine the days when biology was *new*. Overarching histories like this help clarify and separate the regimes and underlying trends, which is useful for understanding the past, predicting the future, and directing our efforts.

Following Harari, I'll use "humans" to refer to the members of the genus homo, and Sapiens to refer to homo Sapiens specifically. I'll use "kya" to refer to "kiloyears ago,"

or thousands of years before now. Harari splits his book into four parts: The Cognitive Revolution, The Agricultural Revolution, The Unification of Humankind, and The Scientific Revolution. I'll discuss those parts separately before giving my thoughts on the whole. Compared to previous reviews I've written for Less Wrong, I'm going to employ quotes far more heavily, primarily because Harari is a good writer who climbs the ladder of abstraction up and down very well, so he often self-summarizes.

The Cognitive Revolution

Harari attributes the cognitive revolution to the ability of Sapiens language (and brains) to communicate about fictions. From an individual perspective, this seems problematic: an individual who only believes true things seems strictly more fit than an individual who believes both true and false things.

But from a collective perspective, fictions can allow cooperation on a much broader scale. This is the first obvious benefit of starting from the lower levels and building up--if you were a Martian biologist, the striking thing about Sapiens is their tremendous ability to cooperate with each other. Other animals do not build cities, and would not find them livable, because there would be too many of their own kind there. But to a city-dwelling human, that humans can meet strangers with only a touch of anxiety seems normal, not necessarily odd enough to demand a powerful explanation.

Many different varieties of animals have tribes, of course, but there doesn't seem to have been much social difference between chimpanzee tribes, elephant tribes, and pre-Cognitive Revolution Sapiens tribes. The social roles of those tribes are often recognizable to us. Typically, the tribe has coalitions that are maintained by close social relationships, frequent contact, and (if the language is developed enough) gossip. But this puts a hard limit on the growth potential of tribes--eventually, the marginal coalition member will get more by joining another coalition than they will from joining the largest coalition, because the largest coalition doesn't have any spare energy, attention, or resources to devote to the potential new member.

But when we go from cooperating based on truth--that we groomed each other yesterday--to cooperating based on fiction--that both of us are intellectual heirs of Bacon and Laplace--the practicalities of group membership change dramatically. We don't have just coalitions supported by pairwise relationships, by principle-oriented factions, where each person in the faction has a relationship primarily with the faction. (Two articles by Yvain, Is Everything a Religion? and I Can Tolerate Everything Except the Outgroup, seem relevant.)

Strictly speaking, of course, the principles underlying the faction do not "exist." As Harari puts it:

> There are no gods in the universe, no nations, no money, no human rights, no laws, and no justice outside the common imagination of human beings.
>
> People easily understand that 'primitives' cement their social order by believing in ghosts and spirits, and gathering each full moon to dance together around the campfire. What we fail to appreciate is that our modern institutions function on exactly the same basis.

But, of course, those modern institutions (as well as the 'primitive' ones) *function*. One division Harari discusses that I found useful was objective, subjective, and inter-

subjective:

> An objective phenomenon exists independently of human consciousness and human beliefs. ... [Radioactivity is his example.]
>
> The subjective is something that exists depending on the consciousness and beliefs of a single individual. ... [A child's imaginary friend is his example.]
>
> The inter-subjective is something that exists within the communication network linking the subjective consciousness of many individuals. If a single individual changes his or her beliefs, or even dies, it is of little importance. However, if most individuals in the network die or change their beliefs, the inter-subjective phenomenon will mutate or disappear. ...
>
> Many of history's most important drivers are inter-subjective: law, money, gods, nations.

That last list looks familiar. Gods and prices are not features of the wavefunction of the physical universe--they're features of communication networks, or cultures. The creation of a third, explicitly defined category (phrases that mean similar things are "social construct" and "myth," at least when used non-pejoratively) solves the epistemic crisis of realizing that many, if not most, of the interesting things in life are *neither* objective nor subjective. The rules of association football are not objective natural laws baked into the universe before there was time, but neither can they be changed by a single person deciding to play differently. (Many authors fall headlong into this epistemic crisis, and Harari every now and then seems to have his presentation, if not his arguments, tripped up by it. But on the whole he manages it well.)

Harari gives the standard history of humans from about 70kya to about 12kya; Sapiens spread out of Africa, decimating and replacing many populations in the way, including both megafauna and other humans. Some managed to contribute genes to modern Sapiens, like Neanderthals, but this is likely cold comfort to animals, peoples, and cultures buried by rivals that were not individually stronger, but more suited to large-scale conflicts.

The Agricultural Revolution

Harari also gives the standard history of the start of agriculture: it seems to have been individually unpleasant but collectively empowering. This is one of the major trends that Harari identifies--the path of history is that stronger collectives absorb or destroy weaker collectives. Much like Dawkins called his account of evolution The Selfish Gene to make obvious that his conception of evolution centered on *genes* instead of on *individuals*, one might call this account of history The Selfish Culture to make obvious that this conception of history centers on *cultures* instead of on *individuals*. (Contrast with Yvain's article centered on individuals.) Harari states this late in the book:

> history's choices are not made for the benefit of humans. There is absolutely no proof that human well-being inevitably improves as history rolls along. There is no proof that cultures that are beneficial to humans must inexorably succeed and spread, while less beneficial cultures must disappear.

And so viewed from Mars, the story of history after agriculture is the story of collectives trading off individual satisfaction for collective power time and time again. Since collective power is what determines survival of a culture, we are left with an immensely strong culture that holds the entire world in its grasp--but, to the individual people living in it, may not actually be any more satisfying than life as a nomad. When we broaden our view to other organisms, domesticated and wild animals make the story even more clear and extreme. Industrially managed cattle are more 'collectively powerful' (in the sense of being economically useful to Sapiens) than wild aurochs herds, but by almost any scale their lives are tremendously miserable compared to their undomesticated ancestors. Sapiens today are not quite industrially managed and only partially domesticated, but potential futures where there is more management and more domestication strike most moderns with horror.

The Unification of Humankind

The ability to reason about inter-subjective objects allows humans to scale their cooperating collectives; larger collectives have more power, intelligence, and investments, which allows them more control over the physical world. This control is reinvested, and the trend of increasing collective power continues. This period of history marks the transition from many independent and mostly unconnected collectives of humans into one connected collective.

> Over the millennia, small simple cultures gradually coalesce into bigger and more complex civilisations, so that the world contains fewer and fewer mega-cultures, each of which is bigger and more complex.

The default behavior of all social animals is to divide the world into "us" and "them," and what is remarkable about Sapiens is the capacity to enlarge the "us" to include more and more power (and individuals). The three trends Harari identifies leading to more unification:

> The first millennium BC witnessed the appearance of three potentially universal orders, whose devotees could for the first time imagine the entire world and the entire human race as a single unit governed by a single set of laws. Everyone was 'us', at least potentially. There was no longer 'them'. The first universal order to appear was economic: the monetary order. The second universal order was political: the imperial order. The third universal order was religious: the order of universal religions such as Buddhism, Christianity, and Islam.
>
> Merchants, conquerors, and prophets were the first people who managed to transcend the binary evolutionary division, 'us vs them', and to foresee the potential unity of humankind. For the merchants, the entire world was a single market and all humans were potential customers. The tried to establish an economic order that would apply to all, everywhere. For the conquerors, the entire world was a single empire and all humans were potential subjects, and for the prophets, the entire world held a single truth and all humans were potential believers. They too tried to establish an order that would be applicable for everyone everywhere.

I won't go through the details of Harari's history of how those three forces worked to bring collectives together. Instead, I'll just share the passages I highlighted from those chapters:

> Christians and Muslims who could not agree on religious beliefs could nevertheless agree on a monetary belief, because whereas religion asks us to believe in something, money asks us to believe that *other people believe in something*.
>
> For thousands of years, philosophers, thinkers, and prophets have besmirched money and called it the root of all evil. Be that as it may, money is also the apogee of human tolerance.
>
> The Chinese Mandate of Heaven was given by Heaven to solve the problems of humankind. The modern Mandate of Heaven will be given by humankind to solve the problems of heaven, such as the hole in the ozone layer and the accumulation of green-house gases. The colour of the global empire may well be green.
>
> When animism was the dominant belief system, human norms and values had to take into consideration the outlook and interests of a multitude of other beings, such as animals, plants, faeries, and ghosts.

(Consider Gwern's article on The Narrowing Circle.)

Harari discusses two philosophical problems that religions suggest answers for: order and evil. It is perhaps surprising that we live in a universe that seems mechanical and orderly, with universal laws closely related to fairly simple mathematics. It is also perhaps surprising that we live in a universe with tremendous amounts of suffering and wickedness.

> So, monotheism explains order, but is mystified by evil. Dualism explains evil, but is puzzled by order. There is one logical way of solving the riddle: to argue that there is a single omnipotent God who created the entire universe - and He's evil. But nobody in history has had the stomach for such a belief.

That no one can stomach it is not quite true, especially if one lets "evil" include indifference. Lovecraft's Cosmicism seems to fit, and Moloch and Gnon both seem to fit. The most heartening version is the "dualism" that sees the mathematics behind the universe as fundamentally uncaring, indifferent, and omnipotent, and humans and human culture as the force of caring, growing in power and strength. One might unite behind the humanist Ahura Mazda against the mathematical Angra Mainyu; one probably won't unite over identification of evolution with Azathoth.

One can see that as an example of what Harari identifies as the dominant religion of the day, "humanism," which he splits into three broad sects:

> Today, the most important humanist sect is liberal humanism, which believes that 'humanity' is a quality of individual humans, and that the liberty of individuals is therefore sacrosanct. According to liberals, the sacred nature of humanity resides within each and every individual Homo sapiens. The inner core of individual humans gives meaning to the world, and is the source for all ethical and political authority. If we encounter an ethical or political dilemma, we should look inside and listen to our inner voice - the voice of humanity. The chief commandments of liberal humanism are meant to protect the liberty of this inner voice against intrusion or harm. These commandments are collectively known as 'human rights'. ...
>
> The liberal belief in the free and sacred nature of each individual is a direct legacy of the traditional Christian belief in the free and eternal souls. Without recourse to

> eternal souls and a Creator God, it becomes embarrassingly difficult for liberals to explain what is so special about individual Sapiens.
>
> Another important sect is socialist humanism. Socialists believe that 'humanity' is a collective rather than individualistic. They hold as sacred not the inner voice of each individual, but the species Homo sapiens as a whole. Whereas liberal humanism seeks as much freedom as possible for individual humans, socialist humanism seeks equality between all humans. According to socialists, inequality is the worst blasphemy against the sanctity of humanity, because it privileges peripheral qualities of humans over their universal essence. For example, when the rich are privileged over the poor, it means that we value money more than the universal essence of all humans, which is the same for rich and poor alike.
>
> Like liberal humanism, socialist humanism is built on monotheist foundations. The idea that all humans are equal is a revamped version of the monotheist conviction that all souls are equal before God. The only humanist sect that has actually broken loose from traditional monotheism is evolutionary humanism, whose most famous representatives are the Nazis. What distinguished the Nazis from other humanist sects was a different definition of 'humanity', one deeply influenced by the theory of evolution. In contrast to the other humanists, the Nazis believed that humankind is not something universal and eternal, but rather a mutable species that can evolve or degenerate. Man can evolve into superman, or degenerate into a subhuman.

The Scientific Revolution

Harari identifies the scientific revolution as a feedback loop between research, power, and resources that led to runaway growth. But 'research' is as old as the cognitive revolution, in the sense that there have always been scholars of one form or another. The three differences underlying the Scientific Revolution were the willingness to admit ignorance, the centrality of observation and mathematics, and the acquisition of new powers.

No rationalist will be surprised by the importance of ignorance, uncertainty, and curiosity. No empiricist will be surprised by the centrality of observation (while the necessity of math remains potentially puzzling). And unless it led to the acquisition of additional powers in the realms of physics, chemistry, or biology, it seems unlikely that the Scientific Revolution would be an epoch of history, rather than just a philosophical curiosity.

> The Scientific Revolution has not been a revolution of knowledge. It has been above all a revolution of ignorance. The great discovery that launched the Scientific Revolution was the discovery that humans do not know the answers to their most important questions.

The description of previous eras centers on what Eliezer calls a "mere Ultimate Prophet." Either an individual person doesn't know something (but could learn it by studying the prophets / ancients), or that thing is unimportant, or it is important and *this new prophet* can explain it. Acceptance of ignorance on the important questions leads to observation on the important questions, and more importantly a sustained community of observations. Instead of taking a single observation and centering a community around that answer ('eating an apple a day will extend your life'), a community can be centered around a *question* ('what behaviors extend lives?').

> Modern-day science is a unique tradition of knowledge, inasmuch as it openly admits *collective* ignorance regarding *the most important questions*. Darwin never argued that he was 'The Seal of the Biologists', and that he had solved the riddle of life once and for all. After centuries of extensive scientific research, biologists admit that they still don't have any good explanation for how brains produce consciousness.
>
> The willingness to admit ignorance has made modern science more dynamic, supple, and inquisitive than any previous tradition of knowledge. This has hugely expanded our capacity to understand how the world works and our ability to invent new technologies.
>
> As science began to solve one unsolvable problem after another, many became convinced that humankind could overcome any and every problem by acquiring and applying new knowledge. Poverty, sickness, wars, famines, old age and death itself were not the inevitable fate of humankind. They were simply the fruits of our ignorance.

A critical part of the feedback loop, of course, is resources. Research is funded by governments, corporations, and individuals because they expect to see some political, economic, or religious good from that research. Progress is not in every direction or a random direction; it is in a direction determined by the values of the funders of research.

The book ends with a description of transhumanism. With power and intelligence comes the ability to design, and historically we have mostly thought about designing our environments. The organisms around us were inherited and partially designed, but we can see a time when farm animals are designed like farm machinery, and even intelligences like ourselves are intelligently designed. But Harari is no techno-optimist who assumes that everything will turn out well; if anything, he draws the trendline towards dystopia. He recognizes the value problem as perhaps the most important issue of the near future.

But, much like Robin Hanson thinking that Ems will lead lives they consider worthwhile, Harari spends about a chapter pointing out that *we* live in a dystopia relative to our ancestors, alienated from many deep relationships that they would have trouble imagining a worthwhile life without. It looks to me like Harari identifies Moloch as a major force of history, and retaining the View From Mars, does not see a point in extolling its virtues or condemning it, asking what *will* happen instead of what *should* happen, and pointing out that we should think long and hard about what should happen, and how to turn that into what will happen.

The other book that I've read that I think this is most similar to is Why The West Rules--for Now by Ian Morris (some discussion on LW here). Both are written by historians, both start their description of the universe 13 billion years ago, and both conclude with transhumanist predictions that the story of the universe from 2050 onwards will differ from the story beforehand in deep and meaningful ways. (Morris makes the claim that the two possibilities are "singularity" and "collapse," with "business as usual"--how most people expect the future to go--being entirely unreasonable on historical grounds.)

But while I just thought Morris's book was neat as an independent summary and confirmation of this view of history, I thought Harari's book was important enough to

write up this article because Sapiens seems to have much crisper abstractions, more developed causal relationships, and more focus on the non-material aspects of historical changes. (It's also shorter, at about 460 pages instead of about 700, which puts it at a length such that I can recommend it more easily.)

Overall, the book is remarkably even-handed and broad, which is difficult when talking about the possibilities of the future, including the values that we might take on. The parts I found weakest were when the I thought the View From Mars slipped, though that may be personal taste (and none of the points seemed egregious enough to quote and attack). I *want* to say it's the best introduction I've come across to this perspective of history and the future and the underlying trends and principles that unite them--but, really, what I mean it's the most compact and readable presentation of them. Whether or not it actually functions as an introduction is something I'm too deep into this view to evaluate (but, once I've subjected my parents to the book, will probably be more able to determine).

Who should read it? I found it interesting, despite having come across most of its component claims before. I suspect that anyone who wants to think deeply about Moloch, values, or the project of socially determining values would benefit from reading the book, if just to see how other branches of thinkers are thinking about these issues. I suspect most people would benefit from reading about early human history at least once, and if you've haven't this is probably a treatment more suited to the interests of LWers than others.

Concept Safety: Producing similar AI-human concept spaces

A frequently-raised worry about AI is that it may reason in ways which are very different from us, and understand the world in a very alien manner. For example, Armstrong, Sandberg & Bostrom (2012) consider the possibility of restricting an AI via "rule-based motivational control" and programming it to follow restrictions like "stay within this lead box here", but they raise worries about the difficulty of rigorously defining "this lead box here". To address this, they go on to consider the possibility of making an AI internalize human concepts via feedback, with the AI being told whether or not some behavior is good or bad and then constructing a corresponding world-model based on that. The authors are however worried that this may fail, because

Humans seem quite adept at constructing the correct generalisations – most of us have correctly deduced what we should/should not be doing in general situations (whether or not we follow those rules). But humans share a common of genetic design, which the OAI would likely not have. Sharing, for instance, derives partially from genetic predisposition to reciprocal altruism: the OAI may not integrate the same concept as a human child would. Though reinforcement learning has a good track record, it is neither a panacea nor a guarantee that the OAIs generalisations agree with ours.

Addressing this, a possibility that I raised in Sotala (2015) was that possibly the concept-learning mechanisms in the human brain are actually relatively simple, and that we could replicate the human concept learning process by replicating those rules. I'll start this post by discussing a closely related hypothesis: that *given a specific learning or reasoning task and a certain kind of data, there is an optimal way to organize the data that will naturally emerge*. If this were the case, then AI and human reasoning might naturally tend to learn the same kinds of concepts, even if they were using very different mechanisms. Later on the post, I will discuss how one might try to verify that similar representations had in fact been learned, and how to set up a system to make them even more similar.

Word embedding

A particularly fascinating branch of recent research relates to the learning of word embeddings, which are mappings of words to very high-dimensional vectors. It turns out that if you train a system on one of several kinds of tasks, such as being able to classify sentences as valid or invalid, this builds up a space of word vectors that reflects the relationships between the words. For example, there seems to be a male/female dimension to words, so that there's a "female vector" that we can add to the word "man" to get "woman" - or, equivalently, which we can subtract from "woman" to get "man". And it so

happens (Mikolov, Yih & Zweig 2013) that we can also get from the word "king" to the word "queen" by adding the same vector to "king". In general, we can (roughly) get to the male/female version of any word vector by adding or subtracting this one difference vector!

Why would this happen? Well, a learner that needs to classify sentences as valid or invalid needs to classify the sentence "the king sat on his throne" as valid while classifying the sentence "the king sat on her throne" as invalid. So including a gender dimension on the built-up representation makes sense.

But gender isn't the only kind of relationship that gets reflected in the geometry of the word space. Here are a few more:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

It turns out (Mikolov et al. 2013) that with the right kind of training mechanism, *a lot* of relationships that we're intuitively aware of become automatically learned and represented in the concept geometry. And like Olah (2014) comments:

> It's important to appreciate that all of these properties of W are *side effects*. We didn't try to have similar words be close together. We didn't try to have analogies encoded with difference vectors. All we tried to do was perform a simple task, like predicting whether a sentence was valid. These properties more or less popped out of the optimization process.

This seems to be a great strength of neural networks: they learn better ways to represent data, automatically. Representing data well, in turn, seems to be essential to success at many machine learning problems. Word embeddings are just a particularly striking example of learning a representation.

It gets even more interesting, for we can use these for translation. Since Olah has already written an excellent exposition of this, I'll just quote him:

> We can learn to embed words from two different languages in a single, shared space. In this case, we learn to embed English and Mandarin Chinese words in the same space.
>
> We train two word embeddings, W*en* and W*zh* in a manner similar to how we did above. However, we know that certain English words and Chinese words have similar meanings. So, we optimize for an additional property: words that we know are close translations should be close together.
>
> Of course, we observe that the words we knew had similar meanings end up close together. Since we optimized for that, it's not surprising. More interesting is that words we *didn't* know were translations end up close together.

> In light of our previous experiences with word embeddings, this may not seem too surprising. Word embeddings pull similar words together, so if an English and Chinese word we know to mean similar things are near each other, their synonyms will also end up near each other. We also know that things like gender differences tend to end up being represented with a constant difference vector. It seems like forcing enough points to line up should force these difference vectors to be the same in both the English and Chinese embeddings. A result of this would be that if we know that two male versions of words translate to each other, we should also get the female words to translate to each other.
>
> Intuitively, it feels a bit like the two languages have a similar 'shape' and that by forcing them to line up at different points, they overlap and other points get pulled into the right positions.

After this, it gets *even more interesting.* Suppose you had this space of word vectors, and then you also had a system which translated *images* into vectors in the same space. If you have images of dogs, you put them near the word vector for dog. If you have images of Clippy you put them near word vector for "paperclip". And so on.

You do that, and then you take some class of images the image-classifier was never trained on, like images of cats. You ask it to place the cat-image somewhere in the vector space. Where does it end up?

You guessed it: in the rough region of the "cat" words. Olah once more:

> This was done by members of the Stanford group with only 8 known classes (and 2 unknown classes). The results are already quite impressive. But with so few known classes, there are very few points to interpolate the relationship between images and semantic space off of.
>
> The Google group did a much larger version – instead of 8 categories, they used 1,000 – around the same time (Frome et al. (2013)) and has followed up with a new variation (Norouzi et al. (2014)). Both are based on a very powerful image classification model (from Krizehvsky et al. (2012)), but embed images into the word embedding space in different ways.
>
> The results are impressive. While they may not get images of unknown classes to the precise vector representing that class, they are able to get to the right neighborhood. So, if you ask it to classify images of unknown classes and the classes are fairly different, it can distinguish between the different classes.
>
> Even though I've never seen a Aesculapian snake or an Armadillo before, if you show me a picture of one and a picture of the other, I can tell you which is which because I have a general idea of what sort of animal is associated with each word. These networks can accomplish the same thing.

These algorithms made no attempt of being biologically realistic in any way. They didn't try classifying data the way the brain does it: they just tried classifying data using whatever worked. And it turned out that this was enough to start constructing a multimodal representation space where a lot of the relationships between entities were similar to the way humans understand the world.

How useful is this?

"Well, that's cool", you might now say. "But those word spaces were constructed from human linguistic data, for the purpose of predicting human sentences. Of course they're going to classify the world in the same way as humans do: they're basically learning the human representation of the world. That doesn't mean that an autonomously learning AI, with its

own learning faculties and systems, is necessarily going to learn a similar internal representation, or to have similar concepts."

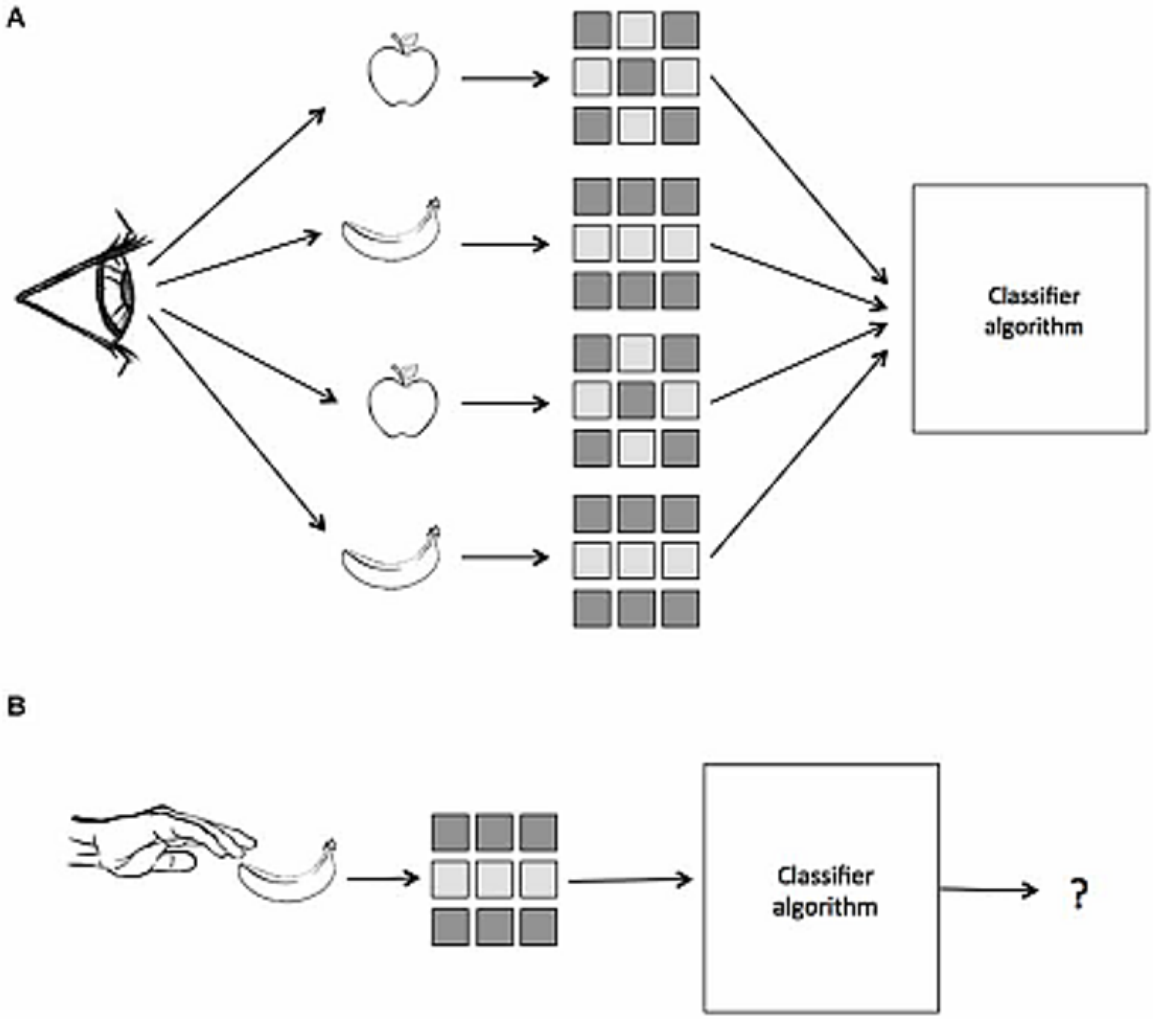
This is a fair criticism. But it is *mildly suggestive* of the possibility that an AI that was trained to understand the world via feedback from human operators would end up building a similar conceptual space. At least assuming that we chose the right learning algorithms.

When we train a language model to classify sentences by labeling some of them as valid and others as invalid, there's a hidden structure implicit in our answers: the structure of how we understand the world, and of how we think of the meaning of words. The language model extracts that hidden structure and begins to classify previously unseen things in terms of those implicit reasoning patterns. Similarly, if we gave an AI feedback about what kinds of actions counted as "leaving the box" and which ones didn't, there would be a certain way of viewing and conceptualizing the world implied by that feedback, one which the AI could learn.

Comparing representations

"Hmm, *maaaaaaaaybe*", is your skeptical answer. "But how would you ever know? Like, you can test the AI in your training situation, but how do you know that it's actually acquired a similar-enough representation and not something wildly off? And it's one thing to look at those vector spaces and *claim* that there are human-like relationships among the different items, but that's still a little hand-wavy. We don't actually *know* that the human brain does anything remotely similar to represent concepts."

Here we turn, for a moment, to neuroscience.



Multivariate Cross-Classification (MVCC) is a clever neuroscience methodology used for figuring out whether different neural representations of the same thing have something in common. For example, we may be interested in whether the visual and tactile representation of a banana have something in common.

We can test this by having several test subjects look at pictures of objects such as apples and bananas while sitting in a brain scanner. We then feed the scans of their brains into a machine learning classifier and teach it to distinguish between the neural activity of looking at an apple, versus the neural activity of looking at a banana. Next we have our test subjects (still sitting in the brain scanners) *touch* some bananas and apples, and ask our machine learning classifier to guess whether the resulting neural activity is the result of touching a banana or an apple. If the classifier - which has *not* been trained on the "touch" representations, only on the "sight" representations - manages to achieve a better-than-chance performance on this latter task, then we can conclude that the neural representation for e.g. "the sight of a banana" has something in common with the neural representation for "the touch of a banana".

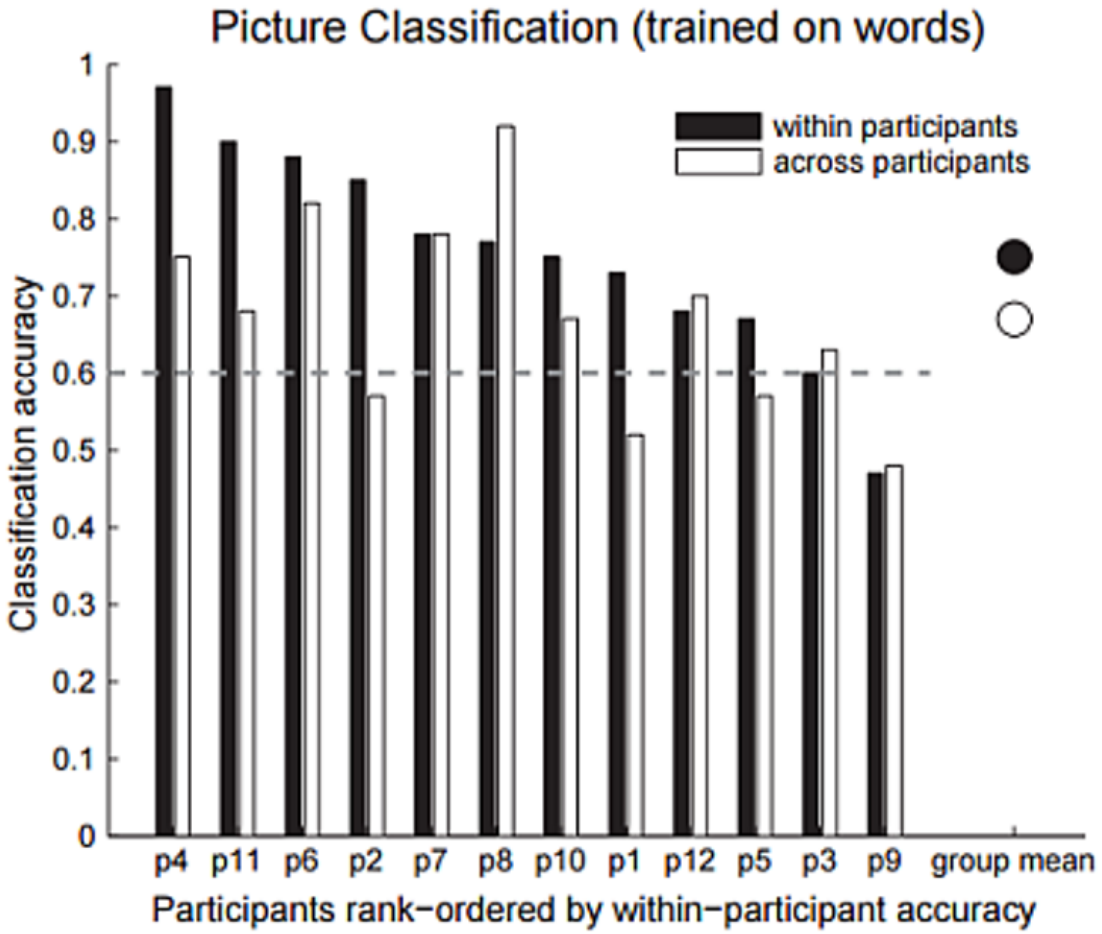


Fig. 2. Classification accuracies across stimulus formats when training on words to identify the category of viewed pictures. Reliable ($p<0.05$) accuracies for identification of category of viewed pictures (filled bars) were reached for 11 participants when training on word data, and reliable ($p<0.05$) accuracies for identification of category of viewed pictures when training on the union of word data from the other participants (unfilled bars) were reached for 8 out of 12 participants. The dashed line indicates the $\alpha = 0.05$ level of significance.

A particularly fascinating experiment of this type is that of Shinkareva et al. (2011), who showed their test subjects both the written words for different tools and dwellings, and, separately, line-drawing images of the same tools and dwellings. A machine-learning classifier was both trained on image-evoked activity and made to predict word-evoked activity and vice versa, and achieved a high accuracy on category classification for both tasks. Even more interestingly, the representations seemed to be similar between subjects. Training the classifier on the word representations of all but one participant, and then having it classify the image representation of the left-out participant, also achieved a reliable ($p<0.05$) category classification for 8 out of 12 participants. This suggests a relatively similar concept space between humans of a similar background.

We can now hypothesize some ways of testing the similarity of the AI's concept space with that of humans. Possibly the most interesting one might be to develop a translation between a human's and an AI's internal representations of concepts. Take a human's neural activation

when they're thinking of some concept, and then take the AI's internal activation when it is thinking of the same concept, and plot them in a shared space similar to the English-Mandarin translation. To what extent do the two concept geometries have similar shapes, allowing one to take a human's neural activation of the word "cat" to find the AI's internal representation of the word "cat"? To the extent that this is possible, one could probably establish that the two share highly similar concept systems.

One could also try to more explicitly optimize for such a similarity. For instance, one could train the AI to make predictions of different concepts, with the additional constraint that its internal representation must be such that a machine-learning classifier trained on a human's neural representations will correctly identify concept-clusters within the AI. This might force internal similarities on the representation beyond the ones that would already be formed from similarities in the data.

Next post in series: The problem of alien concepts.

How to sign up for Alcor cryo

I wrote an article about the process of signing up for cryo since I couldn't find any such accounts online. If you have questions about the sign-up process, just ask.

> A few months ago, I signed up for Alcor's brain-only cryopreservation. The entire process took me 11 weeks from the day I started till the day I received my medical bracelet (the thing that'll let paramedics know that your dead body should be handled by Alcor). I paid them $90 for the application fee. From now on, every year I'll pay $530 for Alcor membership fees, and also pay $275 for my separately purchased life insurance.

http://specterdefied.blogspot.com/2015/04/how-to-sign-up-for-alcor-cryo.html

Replacing guilt

This is a linkpost for https://mindingourway.com/replacing-guilt/

In my experience, many people are motivated primarily by either guilt, shame, or some combination of the two. Some are people who binge-watch television, feel deeply guilty about it, and convert that guilt into a burning need to Actually Do Something on the following day. Others are people who feel guilty whenever they stop working before they *literally fall over from exhaustion,* and in attempts to avoid that guilty feeling, they consistently work themselves weary.

I find that using guilt as a motivation source is both unhealthy and inefficient, but yet, I find it to be a common practice, especially among effective altruists.

Thus, in the coming series of posts, I'm going to explore a whole slew of tools for removing guilt-based motivation and replacing it with something that is both healthier and stronger.

My goal is to help people remove guilt-based motivation entirely, and replace it with intrinsic motivation. I'm aiming to both reduce the frequency of netflix binges *and* reduce the bad feelings that follow. I'm aiming to help people feel like they're still worthwhile human beings if they stop working before they literally drop. I'd like to help people avoid the failure mode where they feel guilty about something for days (even after learning their lesson), and I'm also hoping to remove some shame-based motivation while I'm in the area.

My first goal will be to address the guilt that comes from a feeling of *listlessness*, the vague feeling of guilt that one might get when they play video games all day, or when they turn desperately towards drugs or parties, in attempts to silence the part of themselves that whispers that there must be something else to life.

This sort of guilt cannot be removed by force of will, in most people. The trick to removing this sort of guilt, I think, is to start exploring that feeling that there must be something else to life, that there must be something more to do — and either find something worth working towards, or find that there really isn't actually anything missing. This first sort of listless guilt, I think, comes from someone who wants to find something else to do, and hasn't yet.

Unfortunately, addressing this sort of guilt isn't as easy as just finding a hobby. In my experience, this listless guilt tends to be found in people who have fallen into the nihilistic trap — people who either believe they can't matter, or who believe that no one can matter. It tends to be found in people who believe that humans only ever do what they want, that nothing is truly "better" than anything else, that there is no such thing as altruism, that "morality" is a pleasant lie — that class of beliefs is the class that I will address first, starting with the Allegory of the Stamp Collector.

I'll post the allegory tomorrow. In the interim, I invite you to devise your own tools for removing the listless guilt: the tools that people develop themselves are often more useful to them than the tools they are given.

Failing with abandon

This is a linkpost for https://mindingourway.com/failing-with-abandon/

This is a short public service announcement: *you don't have to fail with abandon.*

Say you're playing Civilization, and your target is to get to sleep before midnight, and you check the clock, and it's already 12:15. If that happens, you *don't* have to say "too late now, I already missed my target" and then keep playing until 4 in the morning.

Say you're trying to eat no more than 2000 calories per day, and then you eat 2300 by the end of dinner, you *don't* have to say "well I already missed my target, so I might as well indulge."

If your goal was to watch only one episode of that one TV show, and you've already watched three, you don't *have* to binge-watch the whole thing.

Over and over, I see people set themselves a target, miss it by a little, and then throw all restraint to the wind. "Well," they seem to think, "willpower has failed me; I might as well over-indulge." I call this pattern "failing with abandon."

But you don't have to fail with abandon. When you miss your targets, you're allowed to say "dang!" and then *continue trying to get as close to your target as you can.*

You don't have to say dang, either. You're allowed to over-indulge, if that's what you want to do. But for lots and lots of people, the idea of missing by *as little as possible* never seems to cross their mind. They miss their targets, and then suddenly they treat their targets as if they were external mandates set by some unjust authority; the jump on the opportunity to defy whatever autarch set an impossible target in the first place; and then (having already missed their target) they reliably fail with abandon.

So this is a public service announcement: you don't have to do that. When you miss your target, you can take a moment to remember who put the target there, and you can ask yourself whether you want to get as close to the target as possible. If you decide you only want to miss your target by a little bit, you still can.

You don't have to fail with abandon.

The stamp collector

This is a linkpost for https://mindingourway.com/the-stamp-collector/

Once upon a time, a group of naïve philosophers found a robot that collected trinkets. Well, more specifically, the robot seemed to collect stamps: if you presented this robot with a choice between various trinkets, it would always choose the option that led towards it having as many stamps as possible in its inventory. It ignored dice, bottle caps, aluminum cans, sticks, twigs, and so on, except insofar as it predicted they could be traded for stamps in the next turn or two. So, of course, the philosophers started calling it the "stamp collector."

Then, one day, the philosophers discovered computers, and deduced out that the robot was merely a software program running on a processor inside the robot's head. The program was too complicated for them to understand, but they did manage to deduce that the robot only had a few sensors (on its eyes and inside its inventory) that it was using to model the world.

One of the philosophers grew confused, and said, "Hey wait a sec, this thing can't be a stamp collector after all. If the robot is only building a model of the world in its head, then it can't be optimizing for its real inventory, because it has no access to its real inventory. It can only ever act according to a model of the world that it reconstructs inside its head!"

"Ah, yes, I see," another philosopher answered. "We did it a disservice by naming it a stamp collector. The robot does not have true access to the world, obviously, as it is only seeing the world through sensors and building a model in its head. Therefore, it must not *actually* be maximizing the number of stamps in its inventory. That would be impossible, because its inventory is outside of its head. Rather, it must be maximizing its *internal stamp counter* inside its head."

So the naïve philosophers nodded, pleased with this, and then they stopped wondering how the stamp collector worked.

There are a number of flaws in this reasoning. First of all, these naïve philosophers have made the homunculus error. The robot's program may not have "true access" to how many stamps were in its inventory (whatever that means), but it *also* didn't have "true access" to its internal stamp counter.

The robot is not occupied by some homunculus that has dominion over the innards but not the outards! The abstract program doesn't have "true" access to the register holding the stamp counter and "fake" access to the inventory. Steering reality towards regions where the inventory has lots of stamps in it is the *same sort of thing* as steering reality towards regions where the stamp-counter-register has high-number-patterns in it. There's not a magic circle containing the memory but not the inventory, within which the robot's homunculus has dominion; the robot program has just as little access to the "true hardware" as it has to the "true stamps."

This brings us to the second flaw in their reasoning reasoning, that of trying to explain choice with a choice-thing. You can't explain why a wall is red by saying "because it's made of tiny red atoms;" this is not an *explanation* of red-ness. In order to explain red-ness, you must explain it in terms of non-red things. And yet, humans have a bad

habit of explaining confusing things in terms of themselves. Why does living flesh respond to mental commands, while dead flesh doesn't? Why, because the living flesh contains Élan Vital. Our naïve philosophers have made the same mistake: they said, "How can it possibly choose outcomes in which the inventory has more stamps? Aha! It must be by choosing outcomes in which the stamp counter is higher!," and in doing so, they have explained choice in terms of choice, rather than in terms of something more basic.

It is *not an explanation* to say "it's trying to get stamps into its inventory because it's trying to maximize its stamp-counter." An explanation would look more like this: the robot's computer runs a program which uses sense-data to build a model of the world. That model of the world contains a representation of how many stapms are in the inventory. The program then iterates over some set of available actions, predicts how many stamps would be in the inventory (according to the model) if it took that action, and outputs the action which leads to the most predicted stamps in its possession.

We could *also* postulate that the robot contains a program which models the world, predicts how the world would change for each action, and *then* predicts how *that* outcome would affect some specific place in internal memory, and *then* selects the action which maximizes the internal counter. That's possible! You could build a machine like that! It's a strictly more complicated hypothesis, and so it gets a complexity penalty, but at least it's an explanation!

And, fortunately for us, it's a *testable* explanation: we can check what the robot does, when faced with the opportunity to directly increase the stamp-counter-register (without actually increasing how many stamps it has). Let's see how that goes over among our naïve philosophers...

> *Hey, check it out: I identified the stamp counter inside the robot's memory. I can't read it, but I did find a way to increase its value. So I gave the robot the following options: take one stamp, or take zero stamps and I'll increase the stamp counter by ten. Guess which one it took?*

"Well, of course, it would choose the latter!" one of the naïve philosophers answers immediately.

> *Nope! It took the former.*

"... Huh! That means that the stampyness of *refusing* to have the stamp counter tampered with must worth be more than 10 stamps!"

> *Huh? What is "stampyness"?*

"Why, stampyness is the robot's internal measure of how much *taking a certain action* would increase its stamp counter."

> *What? That's ridiculous. I'm pretty sure it's just collecting stamps.*

"Impossible! The program doesn't have access to how many stamps it really has; that's a property of the outer world. The robot *must* be optimizing according to values that are actually in its head."

> *Here, let's try offering it the following options: either I'll give it one stamp, or I'll increase its stamp counter by Ackermann(g_{64}, g_{64}) — oh look, it took the stamp.*"

"Wow! That was a very big number, so that almost surely mean that the stampyness of refusing is dependent upon how much stampyness it's refusing! It must be very happy, because you just gave it a *lot* of stampyness by giving it such a compelling offer to refuse."

> *Oh, here, look, I just figured out a way to set the stamp counter to maximum. Here, I'll try offering it a choice between either (a) one stamp, or (b) I'll set the stamp counter to maxi — oh look, it already took the stamp.*

"Incredible! That must there must be some other counter measuring *micro*-stampyness, the amount of stampiness it gets *immediately* upon selecting an action, before you have a chance to modify it! Ah, yes, that's the only possible explanation for why it would refuse you setting the stamp counter to maximum, it *must* be choosing according to the perceived immediate micro-stampyness of each available action! Nice job doing science, my dear fellow, we have learned a lot today!"

Ahh! No! Let's be very clear about this: the robot is predicting which *outcomes* would follow from which actions, and it's ranking them, and it's taking the actions that lead to the best outcomes. Actions are rated according to *what they achieve*. Actions do not *themselves* have intrinsic worth!

Do you see where these naïve philosophers went confused? They have postulated an agent which treats *actions* like *ends*, and tries to steer towards whatever *action* it most prefers — as if actions were ends unto themselves.

You can't explain why the agent takes an action by saying that it ranks actions according to whether or not taking them is good. That begs the question of which actions are good!

This agent rates actions as "good" if they lead to outcomes where the agent has lots of stamps in its inventory. Actions are rated according to what they achieve; they do not themselves have intrinsic worth.

The robot program doesn't contain reality, but it doesn't need to. It still gets to *affect* reality. If its model of the world is correlated with the world, and it takes actions that it predicts leads to more *actual* stamps, then it will tend to accumulate stamps.

It's *not* trying to steer the future towards places where it happens to have selected the most micro-stampy actions; it's just steering the future towards worlds where it predicts it will actually have more stamps.

Now, let me tell you my second story:

Once upon a time, a group of naïve philosophers encountered a group of human beings. The humans seemed to keep selecting the actions that gave them pleasure. Sometimes they ate good food, sometimes they had sex, sometimes they made money to spend on pleasurable things later, but always (for the first few weeks) they took actions that led to pleasure.

But then one day, one of the humans gave lots of money to a charity.

"How can this be?" the philosophers asked, "Humans are pleasure-maximizers!" They thought for a few minutes, and then said, "Ah, it must be that their pleasure from

giving the money to charity outweighed the pleasure they would have gotten from spending the money."

Then a mother jumped in front of a car to save her child.

The naïve philosophers were stunned, until suddenly one of their number said, "I get it! The immediate micro-pleasure of *choosing that action* must have outweighed —

People will tell you that humans always and only ever do what brings them pleasure. People will tell you that there is no such thing as altruism, that people only ever do what they want to.

People will tell you that, because we're trapped inside our heads, we only ever get to care about things inside our heads, such as our own wants and desires.

But I have a message for you: You can, in fact, care about the outer world.

And you can steer it, too. If you want to.