Best of LessWrong: November 2012

Best of LessWrong: November 2012

AI risk-related improvements to the LW wiki

Back in May, Luke suggested the creation of a scholarly AI risk wiki, which was to include a large set of summary articles on topics related to AI risk, mapped out in terms of how they related to the central debates about AI risk. In response, Wei Dai suggested that among other things, the existing Less Wrong wiki could be improved instead. As a result, the Singularity Institute has massively improved the LW wiki, in preparation for a more ambitious scholarly AI risk wiki. The outcome was the creation or dramatic expansion of the following articles:

- 5-and-10
- Acausal Trade
- Acceleration thesis
- Agent
- AGI chaining
- AGI skepticism
- AGI Sputnik moment
- AI advantages
- AI arms race
- AI Boxing
- AI-complete
- AI takeoff
- AIXI
- Algorithmic complexity
- Anvil problem
- Astronomical waste
- Bayesian decision theory
- Benevolence
- Ben Goertzel
- Bias
- Biological Cognitive Enhancement
- Brain-computer interfaces
- Carl Shulman
- Causal decision theory
- Church-Turing thesis
- Coherent Aggregated Volition
- Coherent Blended Volition
- Coherent Extrapolated Volition
- Computing overhang
- Computronium
- Consequentialism
- Counterfactual mugging
- Creating Friendly AI
- Cyc
- Decision theory
- Differential intellectual progress
- Economic consequences of AI and whole brain emulation
- Eliezer Yudkowsky
- Empathic inference
- Emulation argument for human-level AI

- EURISKO
- Event horizon thesis
- Evidential Decision Theory
- Evolutionary algorithm
- Evolutionary argument for human-level AI
- Existential risk
- Expected utility
- Expected value
- Extensibility argument for greater-than-human intelligence
- FAI-complete
- Fallacy
- Fragility_of_value
- Friendly AI
- Fun Theory
- Future of Humanity Institute
- Game theory
- Gödel machine
- Great Filter
- History of AI risk thought
- Human-AGI integration and trade
- Induction
- Infinities in ethics
- Information hazard
- Instrumental convergence thesis
- Intelligence
- Intelligence explosion
- Jeeves Problem
- Lifespan dilemma
- Machine ethics
- Machine learning
- Malthusian Scenarios
- Metaethics
- Moore's law
- Moral divergence
- Moral uncertainty
- Nanny AI
- Nanotechnology
- Neuromorphic AI
- Nick Bostrom
- Nonperson predicate
- Observation selection effect
- Ontological crisis
- Optimal philanthropy
- Optimization power
- Optimization process
- Oracle AI
- Orthogonality thesis
- Paperclip maximizer
- Pascal's mugging
- Prediction market
- Preference
- Prior probability
- Probability theory
- Recursive self-improvement

- Reflective decision theory
- Regulation and AI risk
- Reinforcement learning
- Search space
- Seed AI
- Self Indication Assumption
- Self Sampling Assumption
- Scoring rule
- Simulation argument
- Simulation hypothesis
- Singleton
- Singularitarianism
- Singularity
- Subgoal stomp
- Superintelligence
- Superorganism
- Technological forecasting
- Technological revolution
- Terminal value
- Timeless decision theory
- Tool AI
- Unfriendly AI
- Universal intelligence
- Utility
- Utility extraction
- Utility indifference
- Value extrapolation
- Value learning
- Whole brain emulation
- Wireheading

In managing the project, I focused on content over presentation, so a number of articles still have minor issues such as the grammar and style having room for improvement. It's our hope that, with the largest part of the work already done, the LW community will help improve the articles even further.

Thanks to everyone who worked on these pages: Alex Altair, Adam Bales, Caleb Bell, Costanza Riccioli, Daniel Trenor, João Lourenço, Joshua Fox, Patrick Rhodes, Pedro Chaves, Stuart Armstrong, and Steven Kaas.

Checklist of Rationality Habits

As you may know, the Center for Applied Rationality has run several workshops, each teaching content similar to that in the core sequences, but made more practical, and more into fine-grained habits.

Below is the checklist of rationality habits we have been using in the minicamps' opening session. It was co-written by Eliezer, myself, and a number of others at CFAR. As mentioned below, the goal is not to assess how "rational" you are, but, rather, to develop a personal shopping list of habits to consider developing. We generated it by asking ourselves, not what rationality content it's useful to *understand*, but what rationality-related actions (or thinking habits) it's useful to actually *do*.

I hope you find it useful; I certainly have. Comments and suggestions are most welcome; it remains a work in progress. (It's also available as a pdf.)

This checklist is meant for your personal use so you can have a wish-list of rationality habits, and so that you can see if you're acquiring good habits over the next year—it's not meant to be a way to get a 'how rational are you?' score, but, rather, a way to notice specific habits you might want to develop. For each item, you might ask yourself: did you last use this habit...

- *Never*
- *Today/yesterday*
- *Last week*
- *Last month*
- *Last year*
- *Before the last year*

1. **Reacting to evidence / surprises / arguments you haven't heard before; flagging beliefs for examination.**
    1. When I see something odd - something that doesn't fit with what I'd ordinarily expect, given my other beliefs - I successfully notice, promote it to conscious attention and think "I notice that I am confused" or some equivalent thereof. *(Example: You think that your flight is scheduled to depart on Thursday. On Tuesday, you get an email from Travelocity advising you to prepare for your flight "tomorrow", which seems wrong. Do you successfully raise this anomaly to the level of conscious attention? (Based on the experience of an actual LWer who failed to notice confusion at this point and missed their plane flight.)*

    2. When somebody says something that isn't quite clear enough for me to visualize, I notice this and ask for examples. *(Recent example from Eliezer: A mathematics student said they were studying "stacks". I asked for an example of a stack. They said that the integers could form a stack. I asked for an example of something that was not a stack.) (Recent example from Anna: Cat said that her boyfriend was very competitive. I asked her for an example of "very competitive." She said that when he's driving and the*

person next to him revs their engine, he must be the one to leave the intersection first—and when he’s the passenger he gets mad at the driver when they don’t react similarly.)

3. I notice when my mind is arguing for a side (instead of evaluating which side to choose), and flag this as an error mode. *(Recent example from Anna: Noticed myself explaining to myself why outsourcing my clothes shopping does make sense, rather than evaluating whether to do it.)*

4. I notice my mind flinching away from a thought; and when I notice, I flag that area as requiring more deliberate exploration. *(Recent example from Anna: I have a failure mode where, when I feel socially uncomfortable, I try to make others feel mistaken so that I will feel less vulnerable. Pulling this thought into words required repeated conscious effort, as my mind kept wanting to just drop the subject.)*

5. I consciously attempt to welcome bad news, or at least not push it away. *(Recent example from Eliezer: At a brainstorming session for future Singularity Summits, one issue raised was that we hadn't really been asking for money at previous ones. My brain was offering resistance, so I applied the "bad news is good news" pattern to rephrase this as, "This point doesn't change the fixed amount of money we raised in past years, so it is good news because it implies that we can fix the strategy and do better next year.")*

2. **Questioning and analyzing beliefs (after they come to your attention).**
    1. I notice when I'm not being curious. *(Recent example from Anna: Whenever someone criticizes me, I usually find myself thinking defensively at first, and have to visualize the world in which the criticism is true, and the world in which it's false, to convince myself that I actually want to know. For example, someone criticized us for providing inadequate prior info on what statistics we'd gather for the Rationality Minicamp; and I had to visualize the consequences of [explaining to myself, internally, why I couldn’t have done any better given everything else I had to do], vs. the possible consequences of [visualizing how it might've been done better, so as to update my action-patterns for next time], to snap my brain out of defensive-mode and into should-we-do-that-differently mode.)*

    2. I look for the actual, historical causes of my beliefs, emotions, and habits; and when doing so, I can suppress my mind's search for justifications, or set aside justifications that weren't the actual, historical causes of my thoughts. *(Recent example from Anna: When it turned out that we couldn't rent the Minicamp location I thought I was going to get, I found lots and lots of reasons to blame the person who was supposed to get it; but realized that most of my emotion came from the fear of being blamed myself for a cost overrun.)*

3. I try to think of a concrete example that I can use to follow abstract arguments or proof steps. *(Classic example: Richard Feynman being disturbed that Brazilian physics students didn't know that a "material with an index" meant a material such as water. If someone talks about a proof over all integers, do you try it with the number 17? If your thoughts are circling around your roommate being messy, do you try checking your reasoning against the specifics of a particular occasion when they were messy?)*

4. When I'm trying to distinguish between two (or more) hypotheses using a piece of evidence, I visualize the world where hypothesis #1 holds, and try to consider the prior probability I'd have assigned to the evidence in that world, then visualize the world where hypothesis #2 holds; and see if the evidence seems more likely or more specifically predicted in one world than the other *(Historical example: During the Amanda Knox murder case, after many hours of police interrogation, Amanda Knox turned some cartwheels in her cell. The prosecutor argued that she was celebrating the murder. Would you, confronted with this argument, try to come up with a way to make the same evidence fit her innocence? Or would you first try visualizing an innocent detainee, then a guilty detainee, to ask with what frequency you think such people turn cartwheels during detention, to see if the likelihoods were skewed in one direction or the other?)*

5. I try to consciously assess prior probabilities and compare them to the apparent strength of evidence. *(Recent example from Eliezer: Used it in a conversation about apparent evidence for parapsychology, saying that for this I wanted $p < 0.0001$, like they use in physics, rather than $p < 0.05$, before I started paying attention at all.)*

6. When I encounter evidence that's insufficient to make me "change my mind" (substantially change beliefs/policies), but is still more likely to occur in world X than world Y, I try to update my probabilities at least a little. *(Recent example from Anna: Realized I should somewhat update my beliefs about being a good driver after someone else knocked off my side mirror, even though it was legally and probably actually their fault—even so, the accident is still more likely to occur in worlds where my bad-driver parameter is higher.)*

3. **Handling inner conflicts; when different parts of you are pulling in different directions, you want different things that seem incompatible; responses to stress.**
    1. I notice when I and my brain seem to believe different things (a belief-vs-anticipation divergence), and when this happens I pause and ask which of us is right. *(Recent example from Anna: Jumping off the Stratosphere Hotel in Las Vegas in a wire-guided fall. I knew it was safe based on 40,000 data points of people doing it without significant injury, but to persuade my brain I had to visualize 2 times the population of my college jumping off and surviving. Also, my brain sometimes seems much more pessimistic, especially about social things, than I am, and is almost always wrong.)*

2. When facing a difficult decision, I try to reframe it in a way that will reduce, or at least switch around, the biases that might be influencing it. *(Recent example from Anna's brother: Trying to decide whether to move to Silicon Valley and look for a higher-paying programming job, he tried a reframe to avoid the status quo bias: If he was living in Silicon Valley already, would he accept a $70K pay cut to move to Santa Barbara with his college friends? (Answer: No.))*

3. When facing a difficult decision, I check which considerations are consequentialist - which considerations are actually about future consequences. *(Recent example from Eliezer: I bought a $1400 mattress in my quest for sleep, over the Internet hence much cheaper than the mattress I tried in the store, but non-returnable. When the new mattress didn't seem to work too well once I actually tried sleeping nights on it, this was making me reluctant to spend even more money trying another mattress. I reminded myself that the $1400 was a sunk cost rather than a future consequence, and didn't change the importance and scope of future better sleep at stake (occurring once per day and a large effect size each day).*

4. **What you do when you find your thoughts, or an argument, going in circles or not getting anywhere.**
   1. I try to find a concrete prediction that the different beliefs, or different people, definitely disagree about, just to make sure the disagreement is real/empirical. *(Recent example from Michael Smith: Someone was worried that rationality training might be "fake", and I asked if they could think of a particular prediction they'd make about the results of running the rationality units, that was different from mine, given that it was "fake".)*

   2. I try to come up with an experimental test, whose possible results would either satisfy me (if it's an internal argument) or that my friends can agree on (if it's a group discussion). *(This is how we settled the running argument over what to call the Center for Applied Rationality—Julia went out and tested alternate names on around 120 people.)*

   3. If I find my thoughts circling around a particular word, I try to taboo the word, i.e., think without using that word or any of its synonyms or equivalent concepts. (E.g. wondering whether you're "smart enough", whether your partner is "inconsiderate", or if you're "trying to do the right thing".) *(Recent example from Anna: Advised someone to stop spending so much time wondering if they or other people were justified; was told that they were trying to do the right thing; and asked them to taboo the word 'trying' and talk about how their thought-patterns were actually behaving.)*

5. **Noticing and flagging behaviors (habits, strategies) for review and revision.**
   1. I consciously think about information-value when deciding whether to try something new, or investigate something that I'm doubtful about. *(Recent*

example from Eliezer: Ordering a $20 exercise ball to see if sitting on it would improve my alertness and/or back muscle strain.) (Non-recent example from Eliezer: After several months of procrastination, and due to Anna nagging me about the value of information, finally trying out what happens when I write with a paired partner; and finding that my writing productivity went up by a factor of four, literally, measured in words per day.)

2. I quantify consequences—how often, how long, how intense. (*Recent example from Anna: When we had Julia take on the task of figuring out the Center's name, I worried that a certain person would be offended by not being in control of the loop, and had to consciously evaluate how improbable this was, how little he'd probably be offended, and how short the offense would probably last, to get my brain to stop worrying.) (Plus 3 real cases we've observed in the last year: Someone switching careers is afraid of what a parent will think, and has to consciously evaluate how much emotional pain the parent will experience, for how long before they acclimate, to realize that this shouldn't be a dominant consideration.)*

6. **Revising strategies, forming new habits, implementing new behavior patterns.**
   1. I notice when something is negatively reinforcing a behavior I want to repeat. *(Recent example from Anna: I noticed that every time I hit 'Send' on an email, I was visualizing all the ways the recipient might respond poorly or something else might go wrong, negatively reinforcing the behavior of sending emails. I've (a) stopped doing that (b) installed a habit of smiling each time I hit 'Send' (which provides my brain a jolt of positive reinforcement). This has resulted in strongly reduced procrastination about emails.)*

   2. I talk to my friends or deliberately use other social commitment mechanisms on myself. *(Recent example from Anna: Using grapefruit juice to keep up brain glucose, I had some juice left over when work was done. I looked at Michael Smith and jokingly said, "But if I don't drink this now, it will have been wasted!" to prevent the sunk cost fallacy.) (Example from Eliezer: When I was having trouble getting to sleep, I (a) talked to Anna about the dumb reasoning my brain was using for staying up later, and (b) set up a system with Luke where I put a + in my daily work log every night I showered by my target time for getting to sleep on schedule, and a — every time I didn't.)*

   3. To establish a new habit, I reward my inner pigeon for executing the habit. *(Example from Eliezer: Multiple observers reported a long-term increase in my warmth / niceness several months after... 3 repeats of 4-hour writing sessions during which, in passing, I was rewarded with an M&M (and smiles) each time I complimented someone, i.e., remembered to say out loud a nice thing I thought.) (Recent example from Anna: Yesterday I rewarded myself using a smile and happy gesture for noticing that I was doing a string of low-priority tasks without doing the metacognition for putting the top priorities on top. Noticing a mistake is a good habit, which*

I’ve been training myself to reward, instead of just feeling bad.)

4. I try not to treat myself as if I have magic free will; I try to set up influences (habits, situations, etc.) on the way I behave, not just rely on my will to make it so. *(Example from Alicorn: I avoid learning politicians’ positions on gun control, because I have strong emotional reactions to the subject which I don’t endorse.) (Recent example from Anna: I bribed Carl to get me to write in my journal every night.)*

5. I use the outside view on myself. *(Recent example from Anna: I like to call my parents once per week, but hadn't done it in a couple of weeks. My brain said, "I shouldn't call now because I'm busy today." My other brain replied, "Outside view, is this really an unusually busy day and will we actually be less busy tomorrow?")*

Causal Universes

Followup to: Stuff that Makes Stuff Happen

> Previous meditation: Does the idea that everything is made of causes and effects meaningfully constrain experience? Can you coherently say how reality might look, if our universe did *not* have the kind of structure that appears in a causal model?

I can describe to you at least one famous universe that *didn't* look like it had causal structure, namely the universe of J. K. Rowling's Harry Potter.

You might think that J. K. Rowling's universe doesn't have causal structure because it contains magic - that wizards wave their wands and cast spells, which doesn't make any sense and goes against all science, so J. K. Rowling's universe isn't 'causal'.

In this you would be completely mistaken. The domain of "causality" is just "stuff that makes stuff happen and happens because of other stuff". If Dumbledore waves his wand and therefore a rock floats into the air, that's causality. You don't even have to use words like 'therefore', let alone big fancy phrases like 'causal process', to put something into the lofty-sounding domain of causality. There's causality anywhere there's a noun, a verb, and a subject: 'Dumbledore's wand lifted the rock.' So far as I could tell, there wasn't anything in *Lord of the Rings* that violated causality.

You might worry that J. K. Rowling had made a continuity error, describing a spell working one way in one book, and a different way in a different book. But we could just suppose that the spell had changed over time. If we actually found ourselves in that apparent universe, and saw a spell have two different effects on two different occasions, we would not conclude that our universe was uncomputable, or that it couldn't be made of causes and effects.

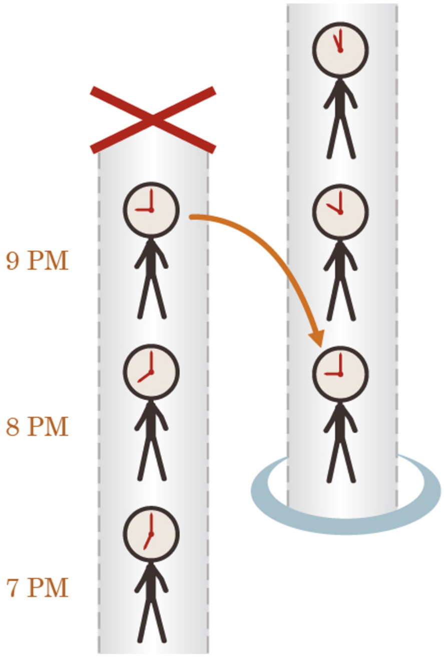
No, the *only* part of J. K. Rowling's universe that violates 'cause and effect' is...

...
...
...

...the Time-Turners, of course.

A Time-Turner, in Rowling's universe, is a small hourglass necklace that sends you back in time 1 hour each time you spin it. In Rowling's universe, this time-travel doesn't allow for *changing* history; whatever you do after you go back, it's already happened. The universe containing the time-travel is a stable, self-consistent object.

If a time machine does allow for changing history, it's easy to imagine how to compute it; you could easily write a computer program which would simulate that universe and its time travel, given sufficient computing power. You would store the state of the universe in RAM and simulate it under the programmed 'laws of physics'. Every nanosecond, say, you'd save a copy of the universe's state to disk. When the Time-Changer was activated at 9pm, you'd retrieve the saved state of the universe from one hour ago at 8pm, load it into RAM, and then insert the Time-Changer and its user in the appropriate place. This would, of course, dump the *rest* of the universe from 9pm into oblivion - no processing would continue onward from that point, which is the same as ending that world and killing everyone in it.[1]



Still, if we don't worry about the ethics or the disk space requirements, then a Time-Changer which can restore and then change the past is easy to *compute.* There's a perfectly clear order of causality in metatime, in the linear time of the simulating computer, even if there are apparent cycles as seen from *within* the universe. The person who suddenly appears with a Time-Changer is the causal descendant of the older universe that just got dumped from RAM.

But what if instead, reality is always - somehow - perfectly self-consistent, so that there's apparently only *one* universe with a future and a past that never changes, so that the person who appears at 8PM has always seemingly descended from *the very same universe* that then develops by 9PM...?

How would you compute *that* in one sweep-through, without any higher-order metatime?

What would a causal graph for *that* look like, when the past descends from its very own future?

And the answer is that there isn't any such causal graph. Causal models are sometimes referred to as DAGs, which stands for Directed Acyclic Graph. If instead there's a *directed cycle*, there's no obvious order in which to compute the joint probability table. Even if you somehow knew that at 8PM somebody was going to appear with a Time-Turner used at 9PM, you still couldn't compute the exact state of the time-traveller without already knowing the future at 9PM, and you couldn't compute the future without knowing the state at 8PM, and you couldn't compute the state at 8PM without knowing the state of the time-traveller who just arrived.

In a causal model, you can compute p(9pm|8pm) and p(8pm|7pm) and it all starts with your unconditional knowledge of p(7pm) or perhaps the Big Bang, but with a

Time-Turner we have p(9pm|8pm) and p(8pm|9pm) and we can't untangle them - multiplying those two conditional matrices together would just yield nonsense.

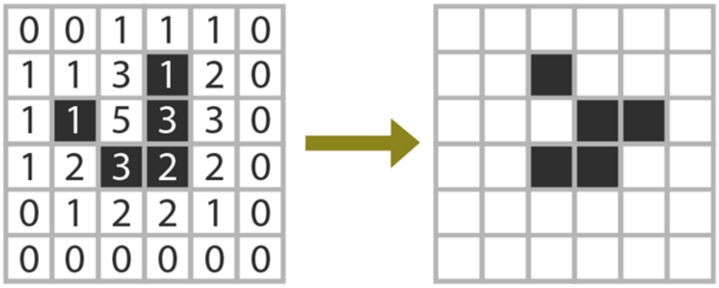
Does this mean that the Time-Turner is beyond all logic and reason?

Complete philosophical panic is basically never justified. We should even be reluctant to say anything like, "The so-called Time-Turner is beyond coherent description; we only think we can imagine it, but really we're just talking nonsense; so we can conclude *a priori* that no such Time-Turner that can exist; in fact, there isn't even a meaningful thing that we've just proven can't exist." This is *also* panic - it's just been made to sound more dignified. The first rule of science is to accept your experimental results, and generalize based on what you see. What if we actually *did* find a Time-Turner that seemed to work like that? We'd just have to accept that Causality As We Previously Knew It had gone out the window, and try to make the best of that.

In fact, despite the somewhat-justified conceptual panic which the protagonist of *Harry Potter and the Methods of Rationality* undergoes upon seeing a Time-Turner, a universe like that can have a straightforward *logical* description even if it has no *causal* description.

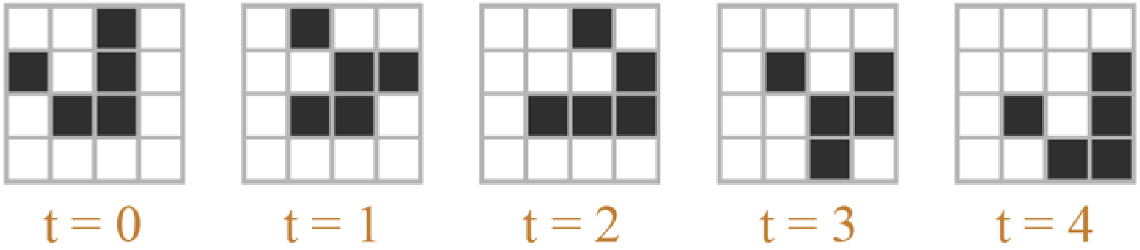
Conway's Game of Life is a very simple specification of a causal universe; what we would today call a cellular automaton. The Game of Life takes place on a two-dimensional square grid, so that each cell is surrounded by eight others, and the Laws of Physics are as follows:

- A cell with 2 living neighbors during the last tick, retains its state from the last tick.
- A cell with 3 living neighbors during the last tick, will be alive during the next tick.
- A cell with fewer than 2 or more than 3 living neighbors during the last tick, will be dead during the next tick.



0	0	1	1	1	0
1	1	3	1	2	0
1	1	5	3	3	0
1	2	3	2	2	0
0	1	2	2	1	0
0	0	0	0	0	0

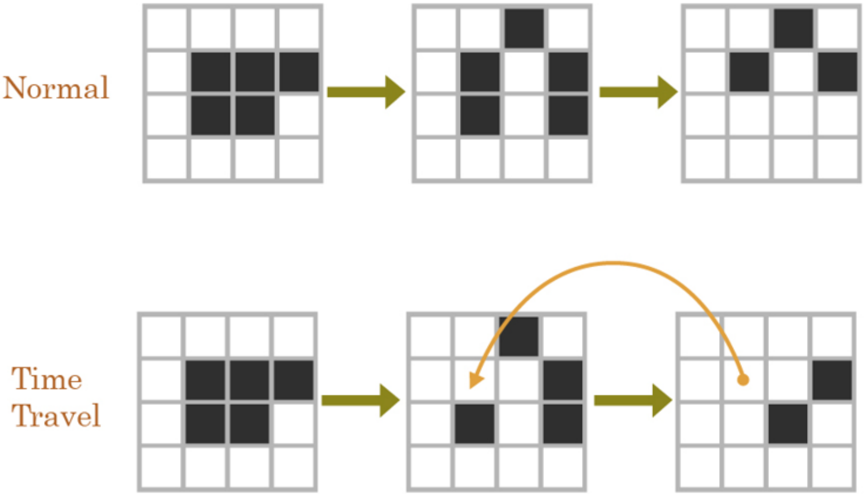
It is my considered opinion that everyone should play around with Conway's Game of Life at some point in their lives, in order to comprehend the notion of 'laws of physics'. Playing around with Life as a kid (on a Mac Plus) helped me gut-level-understand the concept of a 'lawful universe' developing under exceptionless rules.



Now suppose we modify the Game of Life universe by adding some prespecified cases of *time travel* - places where a cell will descend from neighbors in the future, instead of the past.

In particular we shall take a 4x4 Life grid, and arbitrarily hack Conway's rules to say:

- On the 2nd tick, the cell at (2,2) will have its state determined by that cell's state on the 3rd tick, instead of its neighbors on the 1st tick.



It's no longer possible to compute the state of each cell at each time *in a causal order* where we start from known cells and compute their not-yet-known causal descendants. The state of the cells on the 3rd tick, depend on the state of the cells on the 2nd tick, which depends on the state on the 3rd tick.

In fact, the time-travel rule, on the same initial conditions, also permits a live cell to travel back in time, not just a dead cell - this just gives us the "normal" grid! Since you can't compute things in order of cause and effect, even though each local rule is deterministic, the global outcome is not determined.

However, you *could* simulate Life with time travel *merely* by brute-force searching through all possible Life-histories, discarding all histories which disobeyed the laws of Life + time travel. If the entire universe were a 4-by-4 grid, it would take 16 bits to specify a single slice through Time - the universe's state during a single clock tick. If the whole of Time was only 3 ticks long, there would be only 48 bits making up a candidate 'history of the universe' - it would only take 48 bits to completely specify a

History of Time. 2^48 is just 281,474,976,710,656, so with a cluster of 2GHz CPUs it would be quite practical to find, for this *rather tiny* universe, the set of all possible histories that obey the *logical relations* of time travel.

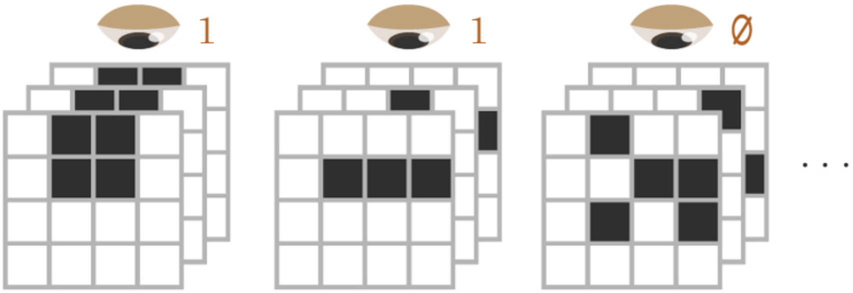
It would no longer be possible to point to a particular cell in a particular history and say, "This is *why* it has the 'alive' state on tick 3". There's no "reason" - in the framework of causal reasons - why the time-traveling cell is 'dead' rather than 'alive', in the history we showed. (Well, except that Alex, in the real universe, happened to pick it out when I asked him to generate an example.) But you could, in principle, find out what the set of permitted histories for a large digital universe, given *lots and lots* of computing power.

Here's an interesting question I do *not* know how to answer: Suppose we had a more *complicated* set of cellular automaton rules, on a vastly larger grid, such that the cellular automaton was large enough, and supported enough complexity, to permit *people* to exist inside it and be computed. Presumably, if we computed out cell states in the ordinary way, each future following from its immediate past, the people inside it would be as real as we humans computed under our own universe's causal physics.

Now suppose that instead of computing the cellular automaton causally, we hack the rules of the automaton to add large time-travel loops - change their physics to allow Time-Turners - and with an *unreasonably large* computer, the size of *two to the power of* the number of bits comprising an *entire history of the cellular automaton,* we enumerate all possible candidates for a universe-history.

So far, we've just generated all 2^N possible bitstrings of size N, for some large N; nothing more. You wouldn't expect this procedure to generate any people or make any experiences real, unless enumerating all finite strings of size N causes all lawless universes encoded in them to be real. There's no causality there, no computation, no law relating one time-slice of a universe to the next...

Now we set the computer to look over this entire set of candidates, and mark with a 1 those that obey the modified relations of the time-traveling cellular automaton, and mark with a 0 those that don't.



If N is large enough - if the size of the possible universe and its duration is large enough - there would be descriptions of universes which experienced natural selection, evolution, perhaps the evolution of intelligence, and of course, time travel with self-consistent Time-Turners, obeying the modified relations of the cellular automaton. And the checker would mark those descriptions with a 1, and all others with a 0.

Suppose we pick out one of the histories marked with a 1 and look at it. It seems to contain a description of people who remember experiencing time travel.

Now, were their experiences real? Did we *make* them real by marking them with a 1 - by applying the logical filter using a causal computer? Even though there was no way of computing future events from past events; even though their universe isn't a causal universe; even though they will have had experiences that literally were not 'caused', that did not have any causal graph behind them, within the framework of their own universe and its rules?

I don't know. *But...*

Our *own* universe does *not* appear to have Time-Turners, and *does* appear to have strictly local causality in which each variable can be computed strictly forward-in-time.

And I don't know *why* that's the case; but it's a likely-looking *hint* for anyone wondering what sort of universes can be real in the first place.

The collection of hypothetical mathematical thingies that can be *described logically* (in terms of relational rules with consistent solutions) looks *vastly* larger than the collection of *causal universes* with locally determined, acyclically ordered events. Most mathematical objects aren't like that. When you say, "We live in a causal universe", a universe that can be computed in-order using local and directional rules of determination, you're *vastly narrowing down the possibilities* relative to all of Math-space.

So it's rather *suggestive* that we find ourselves in a causal universe rather than a logical universe - it suggests that not all mathematical objects can be real, and the sort of thingies that *can* be real and have people in them are constrained to somewhere in the vicinity of 'causal universes'. That you can't have consciousness without computing an agent made of causes and effects, or maybe something can't be real at all unless it's a fabric of cause and effect. It suggests that if there *is* a Tegmark Level IV multiverse, it isn't "all logical universes" but "all causal universes".

Of course you also have to be a bit careful when you start assuming things like "Only causal things can be real" because it's so easy for Reality to come back at you and shout "WRONG!" Suppose you thought reality had to be a *discrete* causal graph, with a finite number of nodes and discrete descendants, *exactly* like Pearl-standard causal models. There would be *no hypothesis in your hypothesis-space* to describe the standard model of physics, where space is continuous, indefinitely divisible, and has complex amplitude assignments over uncountable cardinalities of points.

Reality is primary, saith the wise old masters of science. The first rule of science is just to go with what you see, and try to understand it; rather than standing on your assumptions, and trying to argue with reality.

But *even so,* it's *interesting* that the pure, ideal structure of causal models, invented by statisticians to reify the idea of 'causality' as simply as possible, looks *much* more like the modern view of physics than does the old Newtonian ideal.

If you believed in Newtonian billiard balls bouncing around, and somebody asked you what sort of things can be real, you'd probably start talking about 'objects', like the billiard balls, and 'properties' of the objects, like their location and velocity, and how the location 'changes' between one 'time' and another, and so on.

But suppose you'd never heard of atoms or velocities or this 'time' stuff - just the causal diagrams and causal models invented by statisticians to represent the simplest possible cases of cause and effect. Like this:

And then someone says to you, "Invent a *continuous analogue* of this."

You wouldn't invent billiard balls. There's no billiard balls in a causal diagram.

You wouldn't invent a single time sweeping through the universe. There's no sweeping time in a causal diagram.

You'd stare a bit at B, C, and D which are the sole nodes determining A, screening off the rest of the graph, and say to yourself:

"Okay, how can I invent a *continuous* analogue of there being three nodes that screen off the rest of the graph? How do I do that with a continuous neighborhood of points, instead of three nodes?"

You'd stare at E determining D determining A, and ask yourself:

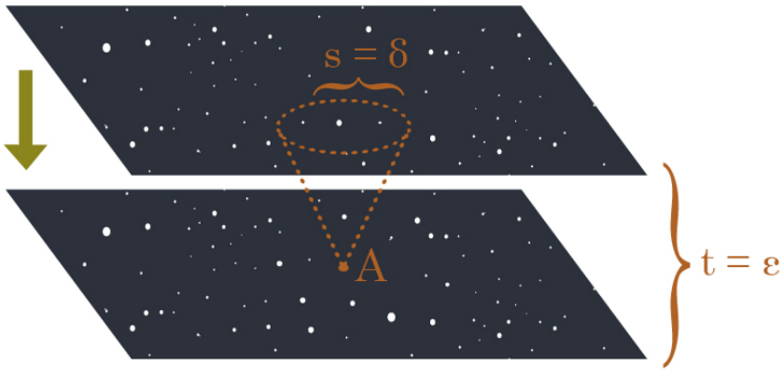
"How can I invent a *continuous* analogue of 'determination', so that instead of E determining D determinining A, there's a continuum of determined points between E and A?"

If you generalized in a certain simple and obvious fashion...

The continuum of relatedness from B to C to D would be what we call *space.*

The continuum of determination from E to D to A would be what we call *time.*

There would be a rule stating that for epsilon time before A, there's a neighborhood of spatial points delta which screens off the rest of the universe from being relevant to A (so long as no descendants of A are observed); and that epsilon and delta can both get arbitrarily close to zero.



There might be - if you were just picking the simplest rules you could manage - a physical constant which related the metric of relatedness (space) to the metric of determination (time) and so enforced a simple continuous analogue of local causality...

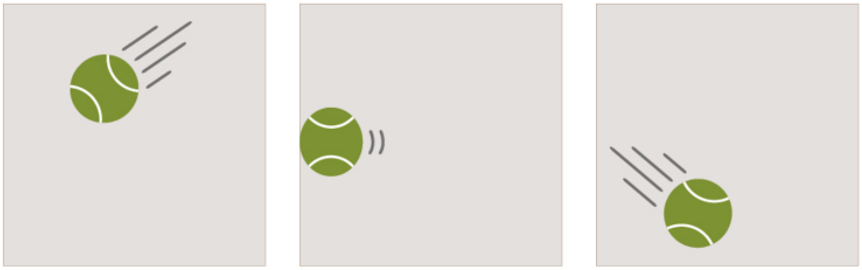
...in our universe, we call it *c,* the speed of light.

And it's worth remembering that Isaac Newton did *not* expect that rule to be there.

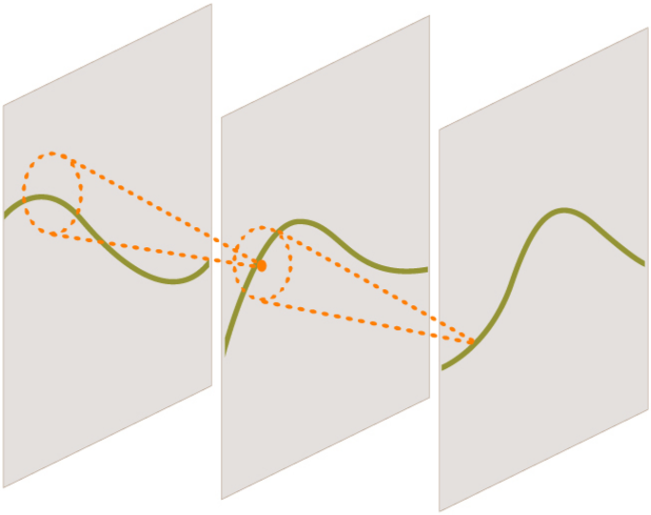
If we just stuck with Special Relativity, and didn't get any *more* modern than that, there would still be little billiard balls like electrons, occupying some particular point in that neighborhood of space.

But if your little neighborhoods of space have billiard balls with velocities, many of which are *slower* than lightspeed... well, that doesn't look like the simplest continuous analogues of a causal diagram, does it?

When we make the first quantum leap and describe particles as waves, we find that the billiard balls have been eliminated. There's no 'particles' with a single point position and a velocity slower than light. There's an electron *field,* and waves propagate through the electron field through points interacting only with locally neighboring points. If a particular electron seems to be moving slower than light, that's just because - even though causality always propagates at exactly *c* between points within the electron *field* - the crest of the electron *wave* can appear to move slower than that. A billiard ball moving through space over time, has been replaced by a set of points with values determined by their immediate historical neighborhood.



vs.



And when we make the second quantum leap into configuration space, we find a timeless universal wavefunction with complex amplitudes assigned over the points in that configuration space, and the amplitude of every point causally determined by its immediate neighborhood in the configuration space.[2]

So, yes, Reality can poke you in the nose if you decide that only discrete causal graphs can be real, or something silly like that.

But on the other hand, taking advice from the math of causality wouldn't always lead you astray. Modern physics looks a *heck* of a lot more similar to "Let's build a continuous analogue of the simplest diagrams statisticians invented to describe theoretical causality", than like anything Newton or Aristotle imagined by looking at the apparent world of boulders and planets.

I don't know what it means... but perhaps we shouldn't ignore the *hint* we received by virtue of finding ourselves inside the narrow space of "causal universes" - rather than the much wider space "all logical universes" - when it comes to guessing what sort of thingies can be real. To the extent we allow non-causal universes in our hypothesis space, there's a strong chance that we are broadening our imagination beyond what can *really* be real under the Actual Rules - whatever *they* are! (It *is* possible to broaden your metaphysics too much, as well as too little. For example, you could allow logical contradictions into your hypothesis space - collections of axioms with no models - and ask whether we lived in one of those.)

If we trusted absolutely that only causal universes could be real, then it would be safe to allow only causal universes into our hypothesis space, and assign probability literally zero to everything else.

But if you were scared of being wrong, then assigning probability literally *zero* means you can't change your mind, ever, even if Professor McGonagall shows up with a Time-Turner tomorrow.

Meditation: Suppose you needed to assign non-zero probability to any way things could conceivably turn out to be, given humanity's rather young and confused state - enumerate all the hypotheses a superintelligent AI should ever be able to arrive at, based on any sort of strange world it might find by observation of Time-Turners or stranger things. How would you enumerate the hypothesis space of all the worlds we could remotely maybe possibly be living in, including worlds with hypercomputers and Stable Time Loops and even stranger features?

Mainstream status.

---

[1] Sometimes I still marvel about how in most time-travel stories nobody thinks of this. I guess it really is true that only people who are sensitized to 'thinking about existential risk' *even notice* when a world ends, or when billions of people are extinguished and replaced by slightly different versions of themselves. But then almost nobody will notice that sort of thing inside their fiction if the characters all act like it's okay.)

[2] Unless you believe in 'collapse' interpretations of quantum mechanics where Bell's Theorem mathematically requires that *either* your causal models don't obey the Markov condition *or* they have faster-than-light nonlocal influences. (Despite a large literature of obscurantist verbal words intended to obscure this fact, as generated and consumed by physicists who don't know about formal definitions of causality or the Markov condition.) If you believe in a collapse postulate, this whole post goes out the window. But frankly, if you believe that, you are bad and you should feel bad.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Mixed Reference: The Great Reductionist Project"

Previous post: "Logical Pinpointing"

Logical Pinpointing

Followup to: Causal Reference, Proofs, Implications and Models

> The fact that one apple added to one apple invariably gives two apples helps in the teaching of arithmetic, but has no bearing on the truth of the proposition that 1 + 1 = 2.
>
> -- James R. Newman, The World of Mathematics

Previous meditation 1: If we can only meaningfully talk about parts of the universe that can be pinned down by chains of cause and effect, where do we find the fact that 2 + 2 = 4? Or did I just make a meaningless noise, there? Or if you claim that "2 + 2 = 4"*isn't* meaningful or true, then what alternate property does the sentence "2 + 2 = 4" have which makes it so much more useful than the sentence "2 + 2 = 3"?

Previous meditation 2: It has been claimed that logic and mathematics is the study of which conclusions follow from which premises. But when we say that 2 + 2 = 4, are we really just *assuming* that? It seems like 2 + 2 = 4 was true well before anyone was around to assume it, that two apples equalled two apples before there was anyone to count them, and that we couldn't make it 5 just by assuming differently.

Speaking conventional English, we'd say the sentence 2 + 2 = 4 is "true", and anyone who put down "false" instead on a math-test would be marked wrong by the schoolteacher (and not without justice).

But what can *make* such a belief true, what is the belief *about,* what is the truth-condition of the belief which can make it true or alternatively false? The sentence '2 + 2 = 4' is true if and only if... what?

In the previous post I asserted that the study of logic is the study of which conclusions follow from which premises; and that although this sort of inevitable implication is sometimes called "true", it could more specifically be called "valid", since checking for inevitability seems quite different from comparing a belief to our own universe. And you could claim, accordingly, that "2 + 2 = 4" is 'valid' because it is an inevitable implication of the axioms of Peano Arithmetic.

And yet thinking about 2 + 2 = 4 doesn't really *feel* that way. Figuring out facts about the natural numbers doesn't feel like the operation of making up assumptions and then deducing conclusions from them. It feels like the numbers are just *out* there, and the only point of making up the axioms of Peano Arithmetic was to *allow* mathematicians to talk about them. The Peano axioms might have been convenient for *deducing* a set of theorems like 2 + 2 = 4, but really all of those theorems were true *about* numbers to begin with. Just like "The sky is blue" is true about the sky, regardless of whether it follows from any particular assumptions.

So comparison-to-a-standard does seem to be at work, just as with *physical* truth... and yet this notion of 2 + 2 = 4 seems different from "stuff that makes stuff happen". Numbers don't occupy space or time, they don't arrive in any order of cause and effect, there are no *events* in numberland.

Meditation: What are we talking *about* when we talk about numbers? We can't navigate to them by following causal connections - so how do we get there from here?

...
...
...

"Well," says the mathematical logician, "that's indeed a very important and interesting question - where are the numbers - but first, I have a question for you. *What* are these 'numbers' that you're talking about? I don't believe I've heard that word before."

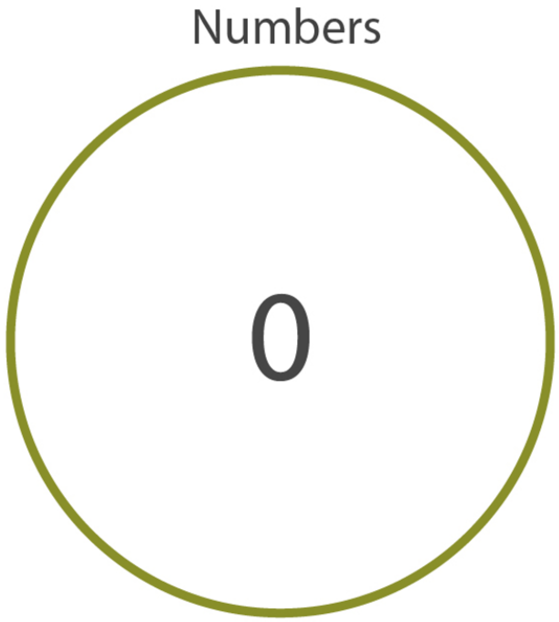
Yes you have.

"No, I haven't. I'm not a typical mathematical logician; I was just created five minutes ago for the purposes of this conversation. So I genuinely don't know what numbers are."

But... you know, 0, 1, 2, 3...

"I don't recognize that 0 thingy - what is it? I'm not asking you to give an exact definition, I'm just trying to figure out what the heck you're talking about in the first place."

Um... okay... look, can I start by asking you to just take on faith that there are these thingies called 'numbers' and 0 is one of them?



"Of course! 0 is a number. I'm happy to believe that. Just to check that I understand correctly, that does mean there exists a number, right?"

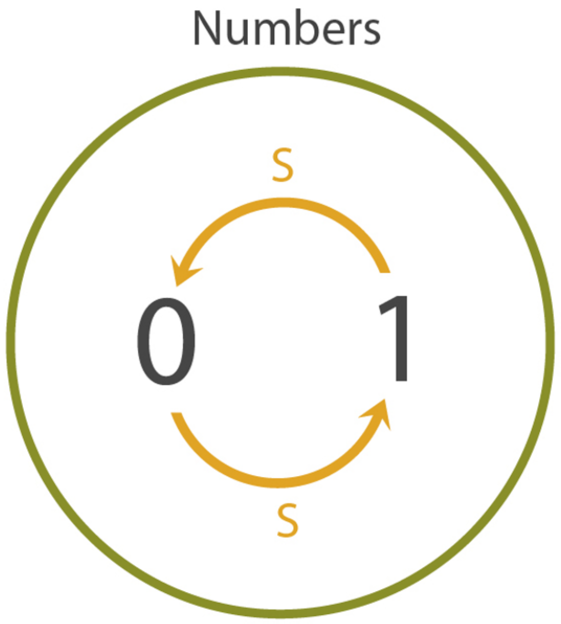
Um, yes. And then I'll ask you to believe that we can take the successor of any number. So we can talk about the successor of 0, the successor of the successor of 0, and so on. Now 1 is the successor of 0, 2 is the successor of 1, 3 is the successor of 2, and so on indefinitely, because we can take the successor of any number -

"In other words, the successor of any number is also a number."

Exactly.

"And in a simple case - I'm just trying to visualize how things might work - we would have 2 equal to 0."

What? No, why would that be -



"I was visualizing a case where there were two numbers that were the successors of each other, so SS0 = 0. I mean, I could've visualized one number that was the successor of itself, but I didn't want to make things *too* trivial -"

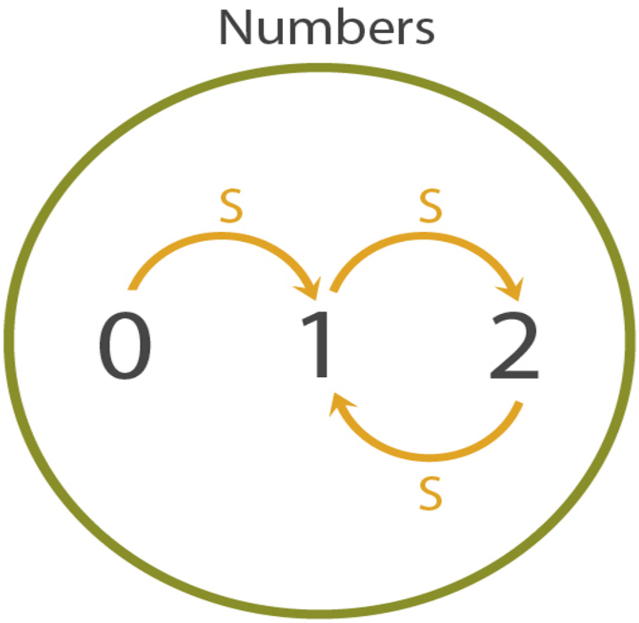
No! That model you just drew - that's *not* a model of the numbers.

"Why not? I mean, what property do the numbers have that this model doesn't?"

Because, um... zero is not the successor of *any* number. Your model has a successor link from 1 to 0, and that's not allowed.

"I see! So we can't have SS0=0. But we could still have SSS0=S0."

What? How -



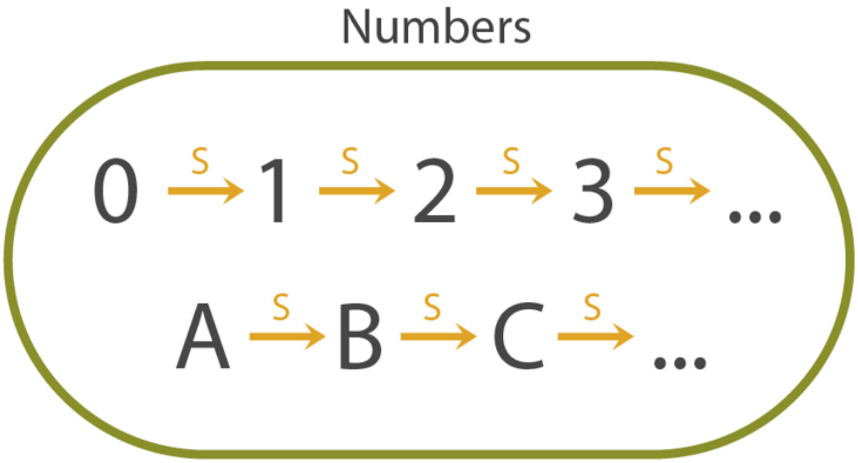
No! Because -

(consults textbook)

- if two numbers have the same successor, they are the same number, that's why! You can't have 2 and 0 *both* having 1 as a successor unless they're the same number, and if 2 was the same number as 0, then 1's successor would be 0, and that's not allowed! Because 0 is not the successor of any number!

"I see. Oh, wow, there's an awful lot of numbers, then. The first chain goes on *forever*."

It sounds like you're starting to get what I - wait. Hold on. What do you mean, the *first* chain -

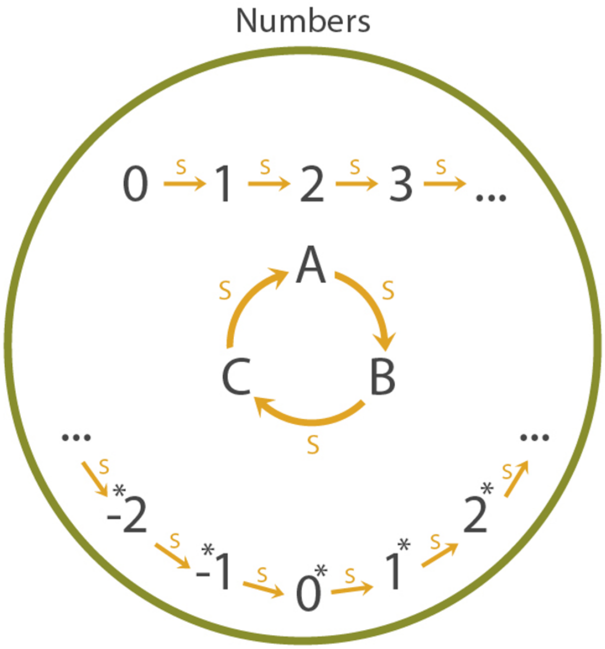


"I mean, you said that there was at least one start of an infinite chain, called 0, but -"

I misspoke. Zero is the *only* number which is not the successor of any number.

"I see, so any other chains would either have to loop or go on forever in *both* directions."

Wha?



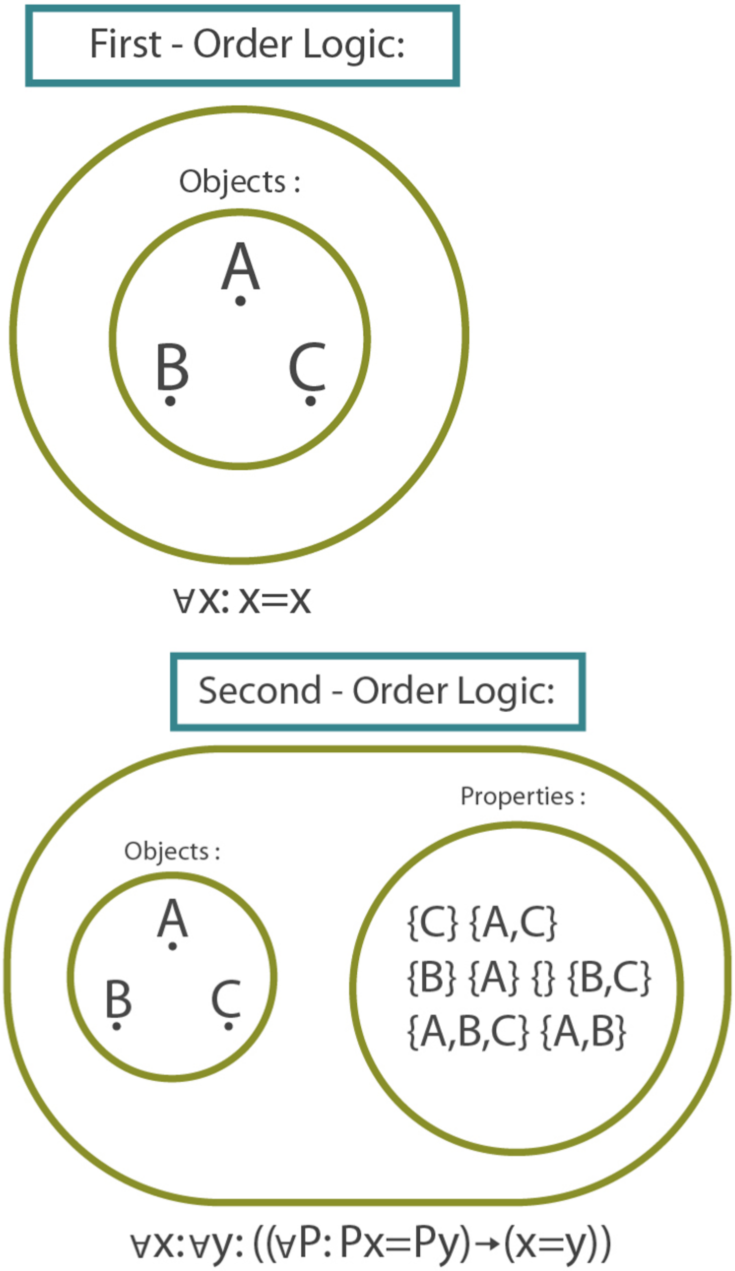
"You said that zero is the only number which is not the successor of any number, that the successor of every number is a number, and that if two numbers have the same successor they are the same number. So, following those rules, any successor-chains besides the one that start at 0 have to loop or go on forever in both directions -"

There *aren't supposed to be any chains* besides the one that starts at 0! Argh! And now you're going to ask me how to say that there shouldn't be any other chains, and I'm not a mathematician so I can't figure out exactly how to -

"Hold on! Calm down. *I'm* a mathematician, after all, so I can help you out. Like I said, I'm not trying to torment you here, just understand what you *mean*. You're right that it's not trivial to formalize your statement that there's only one successor-chain in the model. In fact, you can't say that *at all* inside what's called *first-order logic.* You have to jump to something called *second-order logic* that has some remarkably different properties (ha ha!) and make the statement there."

What the heck is second-order logic?

"It's the logic of properties! First-order logic lets you quantify over *all objects* - you can say that all objects are red, or all objects are blue, or '∀x: red(x)→¬blue(x)', and so on. Now, that 'red' and 'blue' we were just talking about - those are *properties,* functions which, applied to any object, yield either 'true' or 'false'. A property divides all objects into two classes, a class inside the property and a complementary class outside the property. So everything in the universe is either blue or not-blue, red or not-red, and so on. And then second-order logic lets you quantify over properties - instead of looking at particular objects and asking whether they're blue or red, we can talk *about* properties in general - quantify over *all possible* ways of sorting the objects in the universe into classes. We can say, 'For all properties P', not just, 'For all objects X'."

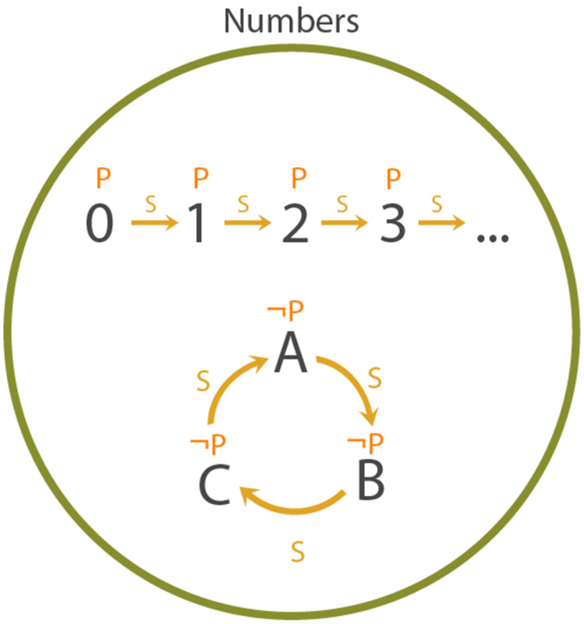


Okay, but what does that have to do with saying that there's only one chain of successors?

"To say that there's only one chain, you have to make the jump to second-order logic, and say that *for all properties P*, if P being true of a number implies P being true of the successor of that number, *and* P is true of 0, *then* P is true of all numbers."

Um... huh. That does sound reminiscent of something I remember hearing about Peano Arithmetic. But how does that solve the problem with chains of successors?

"Because if you had another *separated* chain, you could have a property P that was true all along the 0-chain, but false along the separated chain. And then P would be true of 0, true of the successor of any number of which it was true, and *not* true of all numbers."



I... huh. That's pretty neat, actually. You thought of that pretty fast, for somebody who's never heard of numbers.

"Thank you! I'm an imaginary fictionalized representation of a very *fast* mathematical reasoner."

Anyway, the next thing I want to talk about is addition. First, suppose that for every x, x + 0 = x. Next suppose that if x + y = z, then x + Sy = Sz -

"There's no need for that. We're done."

What do you mean, we're done?

"Every number has a successor. If two numbers have the same successor, they are the same number. There's a number 0, which is the only number that is not the successor of any other number. And every property true at 0, and for which P(Sx) is true whenever P(x) is true, is true of all numbers. In combination, those premises narrow down a *single* model in mathematical space, up to isomorphism. If you show me two models matching these requirements, I can perfectly map the objects and successor relations in them. You can't add any new object to the model, or subtract an object, without violating the axioms you've already given me. It's a uniquely identified mathematical collection, the objects and their structure *completely pinned down*. Ergo, there's no point in adding any more requirements. Any meaningful statement you can make about these 'numbers', as you've defined them, is *already true* or *already false* within that pinpointed model - its truth-value is already semantically implied by the axioms you used to talk about 'numbers' as opposed to something else. If the new axiom is already true, adding it won't change what the previous axioms *semantically* imply."

Whoa. But don't I have to define the + operation before I can talk about it?

"Not in second-order logic, which can quantify over relations as well as properties. You just say: 'For every relation R that works exactly like addition, the following statement Q is true about that relation.' It would look like, '$\forall$ relations R: $(\forall x \forall y \forall z: (R(x, 0, z) \leftrightarrow (x=z)) \land (R(x, Sy, z) \leftrightarrow R(Sx, y, z))) \rightarrow Q)$', where Q says whatever you meant to say about +, using the token R. Oh, sure, it's more convenient to add + to the language, but that's a mere *convenience* - it doesn't change which facts you can prove. Or to say it outside the system: So long as I *know* what numbers are, you can just explain to me how to add them; that doesn't change which mathematical structure we're already talking about."

...Gosh. I think I see the idea now. It's not that 'axioms' are mathematicians asking for you to just assume some things about numbers that seem obvious but can't be proven. Rather, axioms *pin down that we're talking about numbers as opposed to something else.*

"Exactly. That's why the *mathematical* study of numbers is *equivalent* to the *logical* study of which conclusions follow inevitably from the number-axioms. When you formalize logic into syntax, and prove theorems like '2 + 2 = 4' by syntactically deriving new sentences from the axioms, you can safely infer that 2 + 2 = 4 is semantically implied within the mathematical universe that the axioms pin down. And there's no way to try to 'just study the numbers without assuming any axioms', because those axioms are how you can talk about *numbers* as opposed to something else. You can't take for granted that just because your mouth makes a sound 'NUM-burz', it's a meaningful sound. The axioms aren't things you're arbitrarily making up, or assuming for convenience-of-proof, about some pre-existent thing called numbers. You need axioms to pin down a mathematical universe before you can talk *about* it in the first place. The axioms are pinning down what the heck this 'NUM-burz' sound means in the first place - that your mouth is talking about 0, 1, 2, 3, and so on."

Could you also talk about unicorns that way?

"I suppose. Unicorns don't exist in reality - there's nothing in the world that behaves like that - but they could nonetheless be described using a consistent set of axioms, so that it would be *valid* if not quite *true* to say that if a unicorn would be attracted to Bob, then Bob must be a virgin. Some people might dispute whether unicorns *must* be attracted to virgins, but since unicorns aren't real - since we aren't locating them within our universe using a causal reference - they'd just be talking about different models, rather than arguing about the properties of a known, fixed mathematical model. The 'axioms' aren't making questionable guesses about some real physical unicorn, or even a mathematical unicorn-model that's already been pinpointed; they're just fictional premises that make the word 'unicorn' talk about something inside a story."

But when I put two apples into a bowl, and then put in another two apples, I get four apples back out, regardless of anything I assume or don't assume. I don't need any axioms at all to get four apples back out.

"Well, you do need axioms to talk about *four,* SSSS0, when you say that you got 'four' apples back out. That said, indeed your experienced outcome - what your eyes see - doesn't depend on what axioms you assume. But that's because the apples are behaving like numbers whether you believe in numbers or not!"

The apples are behaving like numbers? What do you mean? I thought numbers were this ethereal mathematical model that got pinpointed by axioms, not by looking at the real world.

"Whenever a part of reality behaves in a way that conforms to the number-axioms - for example, if putting apples into a bowl obeys rules, like no apple spontaneously appearing or vanishing, which yields the high-level behavior of numbers - then all the mathematical theorems we proved valid in the universe of numbers can be imported back into reality. The conclusion isn't absolutely certain, because it's not absolutely certain that nobody will sneak in and steal an apple and change the physical bowl's behavior so that it doesn't match the axioms any more. But so long as the premises are true, the conclusions are true; the conclusion can't fail unless a premise also failed. You get four apples in reality, because those apples *behaving numerically* isn't something you *assume,* it's something that's *physically true.* When two clouds collide and form a bigger cloud, on the other hand, they aren't behaving like integers, whether you assume they are or not."

But if the awesome hidden power of mathematical reasoning is to be imported into parts of reality that behave like math, why not reason about apples in the first place instead of these ethereal 'numbers'?

"Because you can prove once and for all that *in any process which behaves like integers,* 2 thingies + 2 thingies = 4 thingies. You can store this general fact, and recall the resulting prediction, for *many* different places inside reality where physical things behave in accordance with the number-axioms. Moreover, so long as we believe that a calculator behaves like numbers, pressing '2 + 2' on a calculator and getting '4' tells us that 2 + 2 = 4 is true of numbers and then to expect four apples in the bowl. It's not like anything fundamentally different from that is going on when we try to add 2 + 2 inside our own *brains* - all the information we get about these 'logical models' is coming from the observation of physical things that allegedly behave like their axioms, whether it's our neurally-patterned thought processes, or a calculator, or apples in a bowl."

I... think I need to consider this for a while.

"Be my guest! Oh, and if you run out of things to think about from what I've said already -"

Hold on.

"- try pondering this one. Why does 2 + 2 come out the same way each time? Never mind the question of why the *laws of physics* are stable - why is *logic* stable? Of course I can't *imagine* it being any other way, but that's not an explanation."

Are you sure you didn't just degenerate into talking bloody nonsense?

"Of *course* it's bloody nonsense. If I knew a way to think about the question that wasn't bloody nonsense, I would already know the answer."

Meditation for next time:

> HUMANS NEED FANTASY TO BE HUMAN.
>
> "Tooth fairies? Hogfathers? Little—"
>
> YES. AS PRACTICE. YOU HAVE TO START OUT LEARNING TO BELIEVE THE *LITTLE* LIES.
>
> "So we can believe the big ones?"
>
> YES. JUSTICE. MERCY. DUTY. THAT SORT OF THING.
>
> "They're not the same at all!"
>
> YOU THINK SO? THEN TAKE THE UNIVERSE AND GRIND IT DOWN TO THE FINEST POWDER AND SIEVE IT THROUGH THE FINEST SIEVE AND THEN *SHOW* ME ONE ATOM OF JUSTICE, ONE MOLECULE OF MERCY.
>
> - Susan and Death, in *Hogfather* by Terry Pratchett

So far we've talked about two kinds of meaningfulness and two ways that sentences can refer; a way of comparing to physical things found by following pinned-down causal links, and logical reference by comparison to models pinned-down by axioms. Is there anything else that can be meaningfully talked about? Where would you find justice, or mercy?

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Causal Universes"

Previous post: "Proofs, Implications, and Models"

Thoughts on designing policies for oneself

Note: This was originally written in relation to this rather scary comment of lukeprog's on value drift. I'm now less certain that operant conditioning is a significant cause of value drift (leaning towards near/far type explanations), but I decided to share my thoughts on the topic of policy design anyway.

Several years ago, I had a reddit problem. I'd check reddit instead of working on important stuff. The more I browsed the site, the shorter my attention span got. The shorter my attention span got, the harder it was for me to find things that were enjoyable to read. Instead of being rejuvenating, I found reddit to be addictive, unsatisfying, and frustrating. Every time I thought to myself that I really should stop, there was always just one more thing to click on.

So I installed LeechBlock and blocked reddit at all hours. That worked really well... for a while.

Occasionally I wanted to dig up something I remembered seeing on reddit. (This wasn't always bad--in some cases I was looking up something related to stuff I was working on.) I tried a few different policies for dealing with this. All of them basically amounted to inconveniencing myself in some way or another whenever I wanted to dig something up.

After a few weeks, I no longer felt the urge to check reddit compulsively. And after a few months, I hardly even remembered what it was like to be an addict.

However, my inconvenience barriers were still present, and they were, well, inconvenient. It really *was* pretty annoying to make an entry in my notebook describing what I was visiting for and start up a different browser just to check something. I figured I could always turn LeechBlock on again if necessary, so I removed my self-imposed barriers. And slid back in to addiction.

After a while, I got sick of being addicted again and decided to do something about it (again). Interestingly, I forgot my earlier thought that I could just turn LeechBlock on again easily. Instead, thinking about LeechBlock made me feel hopeless because it seemed like it ultimately hadn't worked. But I did try it again, and the entire cycle then finished repeating itself: I got un-addicted, I removed LeechBlock, I got re-addicted.

This may seem like a surprising lack of self-awareness. All I can say is: Every second my brain gathers tons of sensory data and discards the vast majority of it. Narratives like the one you're reading right now don't get constructed on the fly automatically. Maybe if I had been following orthonormal's advice of keeping and monitoring a record of life changes attempted, I would've thought to try something different.

Anyway, what finally worked was setting up a site blocker that blocked the reddit.com homepage only. There was no inconvenience associated with visiting other pages, so the "willpower upkeep cost" of this policy was pretty minimal. I drew a mental "line in the sand" prohibiting me from ever loading a web page just to see what had changed on it (excluding email and some other stuff), and this rough heuristic (which I've safely gotten informal with) has served me well ever since.

The point of this anecdote is: Having well-designed policies matters. In the same way that the laws of a nation or the rules of a board game are very important, the policies you set up for yourself to follow are very important. ("Consequentialism is what's correct; virtue ethics is what works for humans.")

You might be wondering how I read Less Wrong, since it's a web page that changes. Less Wrong is a tough one, because it's got the variable reinforcement that makes reddit addictive, but hanging out here can also be a pretty good use of time. Lately what I've been doing is using Google Reader as one of my go-to break activities, and stuffing it so full of feeds that there's a growing backlog of interesting stuff to read every time I visit. The idea here is to have a constant reinforcer instead of a variable one, and it seems to work as far as avoiding addiction is concerned.

Policy design tips

My reddit experience illustrates a few recommendations for designing policies:

- Make the willpower upkeep cost of your policy as low as possible. The higher the upkeep cost, the greater the chance that you'll lack the willpower to uphold it at some point, and thereby lose the cognitive momentum you've got behind it. Willpower spent on upkeep is also willpower that can't be spent on other stuff. (Yes, I know that the resource of model of willpower isn't perfect, but it seems pretty descriptive in my case and I haven't figured out how to subvert it significantly. If you have, please write a post about it!) I think this is the reason why the "cheat days" in diets like Tim Ferris' work--eating whatever you want one day per week decreases the diet's willpower upkeep from impossible to bearable.
- Don't berate yourself, debug your policies. In the same way having your program work correctly on the first run is not the default, successfully acting like an agent on the first try is not the default. Just like in programming, you should *expect* to fix some bugs before getting something that works. (Another way to think about it: You're trying to build a multi-story structure on a planet with extremely high gravity. The high gravity represents the strength of your instincts to just do whatever feels good and doesn't feel bad. Your first few structures fall down, but eventually you manage to figure out a blueprint for a structure that doesn't fall down. Even if falls later, that doesn't matter too much 'cause you can just rebuild it, probably with a few improvements.)
- If possible, have a clear line between policy-compliant and policy-breaking behaviour, to guard against slippery slopes.

The rest of this post is going to consist of more policy design advice. I don't remember the policy design attempts that spawned each piece of advice, and my advice may not work for you. But hopefully it will make for a good starting point for your own policy design experiments.

Consistency

An overarching principle: As much as possible, you want to there to be consistency between what you tell yourself to do and what you actually do. If you've been telling yourself to do something and it's not working, stop. Step back, gain some self-awareness, get creative, and try to figure out some other way to modify your behaviour.

Why is this so important? Because ignoring what you tell yourself to do is a *really bad habit*. Let's say I'm trying to lose weight. Every morning I tell myself that I shouldn't eat a cookie at lunch, and every afternoon I give in and eat a cookie. This amounts to reinforcing the behaviour of rationalizing my way around my diet! The more times I rationalize my way around my diet and get rewarded with a tasty cookie, the stronger my habit of breaking diets is going to become. It might even be a good idea for me to stop trying to diet completely for a while until the behaviour of rationalizing my way around my diet dies off.

I also think the game-theoretic view of time-inconsistency is useful. If you build up a track record of self-cooperation, following pre-commitments becomes easier because you know that by breaking the pre-commitment, you'll be destroying something valuable. Part of this is not making excessive demands on yourself so that track record can actually be built up in the first place. See also: How I Lost 100 Pounds Using TDT.

If you keep these arguments in mind when your brain starts making "just this once" type arguments, hopefully you'll be better at resisting them.

Translating guilt in to policy ideas

- Instead of feeling guilty about something you find yourself doing, think "what policy should I make"? Then continue doing the activity guiltlessly and implement the policy when you're feeling energetic later. (Implementing the policy later means you'll be less likely to break it right after having made it.) Untargeted guilt isn't that useful, and it's especially useless if you're going to do the behaviour anyway. It's much better to translate your vague suggestions for yourself in to a set of specific guidelines, and then put enough momentum behind those guidelines that they don't take much willpower to follow.
- For me, there seems to be a very strong effect where if I make a policy when I'm not feeling very high-willpower, I won't take it seriously and will ignore it later on. So I recommend just noting down policy ideas if you're feeling tired. Then you can refine them and commit to actually following them later on.

Refining policy ideas

(Suggestion: Refer back to this list when you're in a high-energy state and you've got a policy you want to implement.)

- You're encouraged to spend a while on thinking about implementation details if the policy is important to you. The longer you spend thinking about and refining your policy, the more cognitive momentum you'll have behind it.
- Keep willpower upkeep low, as mentioned above.
- Do brainstorming in a text document and finish with a written description of your policy.
- As you think about your policy, brainstorm a to-do list of ways you could modify your environment to make following your policy easier (ex.: throw out your cigarettes, tape reminders on stuff).
- If you don't remember that you made a policy, then violating it probably shouldn't count as a "real" violation. Your memory isn't perfect, why try to affect what you can't control?
- For each policy you make for yourself, I recommend giving yourself two ways to change it. The "slow" way: make a change and it comes in to effect X hours later. The "fast" way: think of a change, think as hard as you can for Y minutes about why it may be a bad change, and if it would seem like a good change to the self that made the policy at the end of Y minutes, make the change. You'll have to decide on X and Y when you make the policy. Policy changes should be reflected in your text document. It may be a good idea to have a "dry run" for a few days with Y = 0 or something like that.
- It also may be a good idea to have two standards for yourself: a carefully-defined "formal" standard, and a higher "informal" standard that isn't as rigorously specified. Try to anchor on your informal standard and follow it in practice, but count it as a win every time you do better than your formal standard. (Think of your formal standard as being at zero, and your informal standard as being at some positive number. Ideally you should have periodic feelings of pride for beating your formal standard, reinforcing the behaviour of following your policy.)

Tips on repairing broken policies

Hopefully this won't actually happen, but let's say you broke your policy. What now?

- Hold the line. Decide that whatever you did to break the policy is allowed for now, but keep the rest of the policy intact.
- Some point later on, when you're feeling energetic, restore your original policy and try to improve it to prevent the particular failure mode you encountered.

I've additionally found regular meditation to be useful for maintaining policies. (Sam Harris on meditation.)

Conclusion

It's been over 6 months since I wrote this article. Here's what my internet distraction policy has evolved in to (it's been stable for the past few months at least, so I thought it might be worth sharing). I have a list of websites that I've classified as "distracting", which include reddit, Less Wrong, and Facebook, but not my email (it's too useful to restrict and I've been able to live with having that one distraction). If I have a reason to visit one of the webpages, I create a log entry in my notebook explaining the reason and then go visit. Sometimes the reason is just "I could use a break right now", and so far using this reason hasn't caused any problems. (If it did, I would probably have to change my policy and hammer out what constitutes a valid reason.) I also open all of the distracting websites on my list in tabs after 11 PM (after a one-minute delay) most days, which means I regularly check my LW/reddit/email inbox and don't have to worry about missing important things in my inbox. For Hacker News in particular, I came up with a more unusual solution: I have a server that's set up to spider the HN homepage every half hour. I originally did this with the intent to write a software tool to browse the homepage archives and filter out all but the best content, but so far I haven't gotten around to this.

Should correlation coefficients be expressed as angles?

Edit 11/28: Edited note at bottom to note that the random variables should have finite variance, and that this is essentially just L^2. Also some formatting changes.

This is something that has been bugging me for a while.

The correlation coefficient between two random variables can be interpreted as the cosine of the angle between them[0]. The higher the correlation, the more "in the same direction" they are. A correlation coefficient of one means they point in exactly the same direction, while -1 means they point in exactly opposite directions. More generally, a positive correlation coefficient means the two random variables make an acute angle, while a negative correlation means they make an obtuse angle. A correlation coefficient of zero means that they are quite literally orthogonal.

Everything I have said above is completely standard. So why aren't correlation coefficients commonly *expressed as angles* instead of as their cosines? It seems to me that this would make them more intuitive to process.

Certainly it would make various statements about them more intuitive. For instance "Even if A is positive correlated with B and B is positively correlated with C, A might be negatively correlated with C." This sounds counterintuitive, until you rephrase it as "Even if A makes an acute angle with B and B makes an acute angle with C, A might make an obtuse angle with C." Similarly, the geometric viewpoint makes it easier to make observations like "If A and B have correlation exceeding $1/\sqrt{2}$ and so do B and C, then A and C are positively correlated" -- because this is just the statement that if A and B make an angle of less than 45° and so do B and C, then A and C make an angle of less than 90°.

Now when further processing is to be done with the correlation coefficients, one wants to leave them as correlation coefficients, rather than take their inverse cosines just to have to take their cosines again later. (I don't know that the angles you get this way are actually useful mathematically, and I suspect they mostly aren't.) My question rather is about when correlation coefficients are *expressed to the reader*, i.e. when they are considered as an end product. It seems to me that expressing them as angles would give people a better intuitive feel for them.

Or am I just entirely off-base here? Statistics, let alone the communication thereof, is not exactly my specialty, so I'd be interested to hear if there's a good reason people don't do this. (Is it assumed that anyone who knows about correlation has the geometric point of view completely down? But most people can't calculate an inverse cosine in their head...)

[0]Formal mathematical version: If we consider real-valued random variables with finite variance on some fixed probability space Ω -- that is to say, $L^2(\Omega)$ -- the covariance is a positive-semidefinite symmetric bilinear form, with kernel equal to the set of essentially constant random variables. If we mod out by these we can consider the result as an inner product space and define angles between vectors as usual, which gives us the inverse cosine of the correlation coefficient. Alternatively we could just take $L^2(\Omega)$ and restrict to those elements with zero mean; this is isomorphic (since

it is the image of the "subtract off the mean" map, whose kernel is precisely the essentially constant random variables).

2012 Less Wrong Census/Survey

11/26: The survey is now closed. Please do not take the survey. Your results will not be counted.

It's that time of year again.

If you are reading this post, and have not been sent here by some sort of conspiracy trying to throw off the survey results, then you are the target population for the Less Wrong Census/Survey. Please take it. Doesn't matter if you don't post much. Doesn't matter if you're a lurker. Take the survey.

This year's census contains a "main survey" that should take about ten or fifteen minutes, as well as a bunch of "extra credit questions". You may do the extra credit questions if you want. **You may skip all the extra credit questions if you want.** They're pretty long and not all of them are very interesting. But it is very important that you not put off doing the survey or not do the survey at all because you're intimidated by the extra credit questions.

The survey will probably remain open for a month or so, but once again **do not delay** taking the survey just for the sake of the extra credit questions.

Please make things easier for my computer and by extension me by reading all the instructions and by answering any text questions in the most obvious possible way. For example, if it asks you "What language do you speak?" please answer "English" instead of "I speak English" or "It's English" or "English since I live in Canada" or "English (US)" or anything else. This will help me sort responses quickly and easily. Likewise, if a question asks for a number, please answer with a number such as "4", rather than "four".

Okay! Enough nitpicky rules! Time to take the...

2012 Less Wrong Census/Survey

Thanks to everyone who suggested questions and ideas for the 2012 Less Wrong Census Survey. I regret I was unable to take all of your suggestions into account, because some of them were contradictory, others were vague, and others would have required me to provide two dozen answers and a thesis paper worth of explanatory text for every question anyone might conceivably misunderstand. But I did make about twenty changes based on the feedback, and *most* of the suggested questions have found their way into the text.

By ancient tradition, if you take the survey you may comment saying you have done so here, and people will upvote you and you will get karma.

LW Women- Minimizing the Inferential Distance

Standard Intro

The following section will be at the top of all posts in the LW Women series.

About two months ago, I put out a call for anonymous submissions by the women on LW, with the idea that I would compile them into some kind of post. There is a LOT of material, so I am breaking them down into more manageable-sized themed posts.

Seven women submitted, totaling about 18 pages.

Crocker's Warning- Submitters were told to not hold back for politeness. You are allowed to disagree, but these are candid comments; if you consider candidness impolite, I suggest you not read this post

To the submittrs- If you would like to respond anonymously to a comment (for example if there is a comment questioning something in your post, and you want to clarify), you can PM your message and I will post it for you. If this happens a lot, I might create a LW_Women sockpuppet account for the submitters to share.

Standard Disclaimer- Women have many different viewpoints, and just because I am acting as an intermediary to allow for anonymous communication does NOT mean that I agree with everything that will be posted in this series. (It would be rather impossible to, since there are some posts arguing opposite sides!)

Please do NOT break anonymity, because it lowers the anonymity of the rest of the submitters.

Minimizing the Inferential Distance

One problem that I think exists in discussions about gender issues between men and women, is that the inferential distance is much greater than either group realizes. Women might assume that men know what experiences women might face, and so not explicitly mention specific examples. Men might assume they know what the women are talking about, but have never really heard specific examples. Or they might assume that these types of things only happened in the past, or not to the types of females in their in-group

So for the first post in this series, I thought it would be worthwhile to try to lower this inferential distance, by sharing specific examples of what it's like as a smart/geeky female. When submitters didn't know what to write, I directed them to this article, by Julia Wise (copied below), and told them to write their own stories. These are not related to LW culture specifically, but rather meant to explain where the women here are coming from. **Warning:** This article is a collection of anecdotes, NOT a logical argument. If you are not interested in anecdotes, don't read it.

Copied from the original article (by a woman on LW) on Radiant Things:

It's lunchtime in fourth grade. I am explaining to Leslie, who has no friends but me, why we should stick together. “We're both rejects,” I tell her. She draws back, affronted. “We're not rejects!” she says. I'm puzzled. It hadn't occurred to me that she wanted to be normal.

....

It's the first week of eighth grade. In a lesson on prehistory, the teacher is trying and failing to pronounce “Australopithecus.” I blurt out the correct pronunciation (which my father taught me in early childhood because he thought it was fun to say). The boy next to me gives me a glare and begins looking for alliterative insults. “Fruity female” is the best he can manage. “Geek girl” seems more apt, but I don't suggest it.

....

It's lunchtime in seventh grade. I'm sitting next to my two best friends, Bridget and Christine, on one side of a cafeteria table. We have been obsessed with Star Wars for a year now, and the school's two male Star Wars fans are seated opposite us. Under Greyson's leadership, we are making up roleplaying characters. I begin describing my character, a space-traveling musician named Anya. “Why are your characters always girls?” Grayson complains. “Just because you're girls doesn't mean your characters have to be.”

“Your characters are always boys,” we retort. He's right, though – female characters are an anomaly in the Star Wars universe. George Lucas (a boy) populated his trilogy with 97% male characters.

....

It's Bridget's thirteenth birthday, and four of us are spending the night at her house. While her parents sleep, we are roleplaying that we have been captured by Imperials and are escaping a detention cell. This is not papers-and-dice roleplaying, but advanced make-believe with lots of pretend blaster battles and dodging behind furniture.

Christine and Cass, aspiring writers, use roleplaying as a way to test out plots in which they make daring raids and die nobly. Bridget, a future lawyer, and I, a future social worker, use it as a way to test out moral principles. Bridget has been trying to persuade us that the Empire is a legitimate government and we shouldn't be trying to overthrow it at all. I've been trying to persuade Amy that shooting stormtroopers is wrong. They are having none of it.

We all like daring escapes, though, so we do plenty of that.

....

It's two weeks after the Columbine shootings, and the local paper has run an editorial denouncing parents who raise "geeks and goths." I write my first-ever letter to the editor, defending geeks as kids parents should be proud of. A girl sidles up to me at the lunch table. "I really liked your letter in the paper," she mutters, and skitters away.

....

It's tenth grade, and I can't bring myself to tell the president of the chess club how desperately I love him. One day I go to chess club just to be near him. There is only one other girl there, and she's really good at chess. I'm not, and I spend the meeting leaning silently on a wall because I can't stand to lose to a boy. Anyway, I despise the girls who join robotics club to be near boys they like, and I don't want to be one of them.

....

It's eleventh grade, and we are gathered after school to play Dungeons and Dragons. (My father, who originally forbid me to play D&D because he had heard it would lead us to hack each other to pieces with axes, has relented.) Christine is Dungeonmaster, and she has recruited two feckless boys to play with us. One of them is in love with her.

(Nugent points out that D&D is essentially combat reworked for physically awkward people, a way of reducing battle to dice rolls and calculations. Christine has been trained by her uncle in the typical swords-and-sorcery style of play, but when she and I play the culture is different. All our adventures feature pauses for our characters to make tea and omelets.)

On this afternoon, our characters are venturing into the countryside and come across two emaciated farmers who tell us their fields are unplowed because dark elves from the forest keep attacking them. “They're going to starve if they don't get a crop in the ground,” I declare. “We've got to plow at least one field.” The boys go along with this plan.

“The farmers tell you their plow has rusted and doesn't work,” the Dungeonmaster informs us from behind her screen.

I persist. “There's got to be something we can use. I look around to see if there's anything else pointy I can use as a plow.”

The Dungeonmaster considers. “There's a metal gate,” she decides.

“Okay, I rig up some kind of harness and hitch it to the pony.”

“It's rusty too,” intones the Dungeonmaster, “and pieces of it keep breaking off. Look, you're not supposed to be farming. You're supposed to go into the forest and find the dark elves. I don't have anything else about the farmers. The elves are the adventure.” Reluctantly, I give up my agricultural rescue plan and we go into the forest to hack at elves.

....

I'm 25 and Jeff's sister's boyfriend is complaining that he never gets to play Magic: the Gathering because he doesn't know anyone who plays. “You could play with Julia,” Jeff suggests.

“Very funny,” says Danner, rolling his eyes.

Jeff and I look at each other. I realize geeks no longer read me as a geek. I still love ideas, love alternate imaginings of how life could be, love being right, but now I care about seeming normal.

“...I wasn't joking,” Jeff says.

“It's okay,” I reassure Danner. “I used to play every day, but I've pretty much forgotten how.”

…………………………

A's Submission

My creepy/danger alert was much higher at a meeting with a high-status (read: supposedly utility-generating, which includes attractive in the sense of pleasing or exciting to look at, but mostly the utility is supposed to be from actions, like work or play) man who was supposed to be my boss for an internship.

The way he talked about the previous intern, a female, the sleazy way he looked while reminiscing and then had to smoke a cigarette, while in a meeting with me, my father (an employer who was abusive), and the internship program director, plus the fact that when I was walking towards the meeting room, the employees of the company, all men, stared at me and remarked, “It’s a girl,” well, I became so creeped out that I didn’t want to go back. It was hard, as a less articulate 16 year-old, to explain to the internship director all that stuff without sounding irrational. But not being able to explain my brain’s priors (incl. abuses that it had previously been too naïve/ignorant to warn against and prevent) wasn’t going to change them or decrease the avoidance-inducing fear and anxiety.

So after some awkward attempts to answer the internship director’s question of why I didn’t want to work there, I asked for a placement with a different company, which she couldn’t do, unfortunately.

B's Submission

Words from my father’s mouth, growing up: “You *need* to be able to cook and keep a clean house, or what man would want to marry you?”

………………

Sixth grade year, I had absolutely no friends whatsoever. A boy I had a bit of a crush on asked me out on a dare. I told him “no,” and he walked back to his laughing friends.

………………

In college I joined the local SCA (medieval) group, and took up heavy weapons combat. The local (almost all-male) “stick jocks” were very supportive and happy to help. Many had even read “The Armored Rose” and so knew about female-specific issues and how to adapt what they were teaching to deal with things like a lower center of gravity, less muscle mass, a different grip, and ingrained cultural hang-ups. The guys were great. But there was one problem: *There was no female-sized loaner armor*.

See, armor is an expensive investment for a new hobby, and so local groups provide loaner armor for newbies, which generally consist of hand-me-downs from the more experienced fighters. We had a decent amount of new female fighters in our college groups, but without a pre-existing generation of female fighters (women hadn’t even been allowed to fight until the 80s) there wasn’t anything to hand down.

The only scar I ever got from heavy combat was armor bite from wearing much-too-large loaner armor. I eventually got my own kit, and (Happy Ending) the upcoming generation of our group always made sure to acquire loaner armor for BOTH genders.

………………

Because of a lack of options, and not really having anywhere else to go, I moved in with my boyfriend and got married at a rather young age (20 and 22, respectively). I had no clue how to be independent. One of the most empowering things I ever did was starting work as an exotic dancer. After years of thinking that I couldn't support myself, it gave me the confidence that I could leave an unhappy marriage without ending up on the street (or more likely, mooching off friends and relatives). Another Happy Ending- Now I'm completely independent.

………………

Walking into the library. A man holds open the door for me. I smile and thank him as I walk through. He makes a sexual comment. I do the Look-Straight-Ahead-and-Walk-Quickly thing.

“Bitch,” he spits out.

It’s not the first of this kind of interaction in my life, and it most *certainly* won’t be the last (almost any time you are in an urban environment, without a male). But it hit harder than most because I had been expecting a polite interaction.

Relevant link: http://goodmenproject.com/ethics-values/why-men-catcall/

………………

The next post will be on Group Attribution Error, and will come out when I get around to it. :P

How minimal is our intelligence?

Gwern suggested that, if it were possible for civilization to have developed when our species had a lower IQ, then we'd still be dealing with the same problems, but we'd have a lower IQ with which to tackle them.   Or, to put it another way, it is unsurprising that living in a civilization has posed problems that our species finds difficult to tackle, because if we were capable of solving such problems easily, we'd probably also have been capable of developing civilization earlier than we did.

How true is that?

In this post I plan to look in detail at the origins of civilization with an eye to considering how much the timing of it did depend directly upon the IQ of our species, rather than upon other factors.

Although we don't have precise IQ test numbers for our immediate ancestral species, the fossil record is good enough to give us a clear idea of how brain size has changed over time:

brain mass as a percent of body mass against time

and we do have archaeological evidence of approximately when various technologies (such as pictograms, or using fire to cook meat) became common.

The First City

About 6,000 years ago (4000 BCE), Ur was a thriving trading village on the flood plain near the mouth of the river Euphrates in what is now called southern Iraq and what historians call Sumeria.

By 3000 BCE it was the heart of a city-state with a core built up populated area covering 37 acres, and would go on over the following thousand years to lead the Sumerian empire, raise a great brick Ziggurat to its patron moon goddess, and become the largest city in the world (65,000 people concentrated in 54 acres).

It was eventually doomed by desertification and soil salination, caused by its own success (over-grazing and land clearing) but, by then, cities had spread throughout the fertile crescent of rivers at the intersection of the European, African and Asian land masses.

Ur may not have been the first city, but it was the first one we know of that wasn't part of a false dawn - one whose culture and technologies did demonstrably spread to other areas. It was the flashpoint.

We don't know for certain what it was about the culture surrounding the dawn of cities that made that particular combination of trade, writing, specialisation, hierarchy and religion communicable, when similar cultures from previous false dawns failed to spread.   We can trace each of those elements to earlier sources, none of them were original to Ur, so perhaps it was a case of a critical mass achieving a self-sustaining reaction.

What we can look at is why the conditions to allow a village to become a large enough city for such a critical mass of developments to accumulate, occurred at that time and place.

From Village to City

Motivation aside, the chief problem with sustaining large numbers of people together in a small area, over several generations, keeping them healthy enough for the population to grow without continual immigration, is ensuring access to a scalable renewable predictable source of calories.

To be predictable means surviving famine years, which requires crops that can be stored for several years, such as grasses (wheat, barley and millet) with large seeds, and good storage facilities to store them in.   It also means surviving pestilence, which requires having a variety of such crops.   To be scalable and renewable means supplying water and nutrients to those crops on an ongoing basis, which requires irrigation and fertiliser from domesticated animals (if you don't have handy regular floods).

Having large mammals available to domesticate, who can provide fertiliser and traction (pulling ploughs and harrows) certainly makes things easier, but doesn't seem to have been a large factor in the timing of the rise of civilisation, or particularly dependent upon the IQ of the human species.   Research suggests that domestication may have been driven as much by the animals own behaviour as by human intention, with those animals daring to approach humans more closely getting first choice of discarded food.

Re-planting seeds to ensure plants to gather in following years, leading to low nutrition grasses adapting into grains with high protein concentrations in the seeds, does seem to a mainly intentional human activity in that we can trace most of the gain in size of such plant species seeds to locations where humans have transitioned from the palaeolithic hunter-gatherer culture (about 2.5 million years ago, to about 10,000 years ago) to the neolithic agricultural culture (about 10,000 year ago, onwards).

Good grain storage seems to have developed incrementally starting with crude stone silo pit designs in 9500 BCE, and progressing by 6000 BCE to customised buildings with raised floors and sealed ceramic containers which could store 80 tons of wheat in good condition for 4 years or more. (Earthenware ceramics date to 25,000 BCE and earlier, though the potter's wheel, useful for mass production of regular storage vessels, does date to the Ubaid period.)

The main key to the timing of the transition from village to city seems to have been not human technology but the confluence of climate and biology. Jared Diamond points the finger at the geography of the region - the fertile crescent farmers had access to a wider variety of grains than anywhere else in the world because that area links and has access to the species of three major land masses.   The Mediterranean climate has a long dry season with a short period of rain, which made it ideal for growing grains (which are much easier to store for several years than, for instance bananas). And everything kicked off when the climate stabilised after the most recent ice age ended about 12,000 years ago.

Ice Ages

Strictly speaking, we're actually talking about the end of a "glacial period" rather than the end of an entire "ice age". The timeline goes:

```
200,000 years ago - 130,000 years ago : glacial period

130,000 years ago - 110,000 years ago : interglacial period

110,000 years ago -  12,000 years ago : glacial period

  12,000 years ago - present : interglacial period
```

So the question now is, why didn't humanity spawn civilisation in the fertile crescent 130,000 years ago, during the last interglacial period? Why did it happen in this one? Did we get significantly brighter in the mean time?

It isn't, on the face of it, an implausible idea. 100,000 years is long enough for evolutionary change to happen, and maybe inventing pottery or becoming farmers did take more brain power than humanity had back then. Or, if not IQ, perhaps it was some other mental change like attention span, or the capacity to obey written laws, live as a specialist in a hierarchy, or similar.

But there's no evidence that this is the case, nor is there a need to hypothesise it because there is at least one genetic change we do know about during that time period, that is by itself sufficient to explain the lack of civilisation 130,000 years ago. And it has nothing to do with the brain.

Brains, Genes and Calories

Using the San Bushpeople as a guide to the palaeolithic diet, hunter-gather culture was able to support an average population density of one person per acre. Not that they ate badly, as individuals. Indeed, they seem to have done better than the early Neolithic farmers. But they had to be free to wander to follow nomadic food sources, and they were limited by access to food that the human body could use to create Docosahexaenoic acid, which is a fatty acid required for human brain development. Originally humans got this from fish living in the lakes and rivers of central Africa. However, about 80,000 years ago, we developed a gene that let us synthesise the same acid from other sources, freeing humanity to migrate away from the wet areas, past the dry northern part, and out into the fertile crescent.

But there is a link between diet and brain. Although the human brain represents only 2% of the body weight, it receives 15% of the cardiac output, 20% of total body oxygen consumption, and 25% of total body glucose utilization. Brains are expensive, in terms of calories consumed. Although brain size or brain activity that uses up glucose is not linearly related to individual IQ, they are linked on a species level.

IQ is polygenetic, meaning that many different genes are relevant to a person's potential maximum IQ. (Note: there are many non-genetic factors that may prevent an individual reaching their potential). Algernon's Law suggests that genes affecting IQ that have multiple alleles still common in the human population are likely to have a cost associated with the alleles tending to increase IQ, otherwise they'd have displaced the competing alleles. In the same way that an animal species that develops the capability to grow a fur coat in response to cold weather is more advanced than one whose genes strictly determine that it will have a thick fur coat at all times, whether the weather is cold or hot; the polygenetic nature of human IQ gives human populations the ability to adapt and react on the time scale of just a few generations, increasing or decreasing the average IQ of the population as the environment changes to reduce or increase the penalties of particular trade-offs for particular alleles contributing to IQ. In particular, if the trade-off for some of those alleles is increased energy consumption and we look at a population of humans moving from an environment where calories are the bottleneck on how many offspring can be produced and survive, to an environment where calories are more easily available, then we might expect to see something similar to the Flynn effect.

Summary

There is no cause to suppose, even if the human genome 100,000 years ago had the full set of IQ-related-alleles present in our genome today, that they would have developed civilisation much sooner.

Comment Navigation Aide

link - DuncanS - animal vs human intelligence
link - DuncanS - brain size & brain efficiency
link - JaySwartz - adaptability vs intelligence
link - RichardKennaway - does more intelligence tend to bring more societal happiness?
link - mrglwrf - Ur vs Uruk
link - NancyLebovitz - does decreased variance of intelligence tend to bring more societal happiness?
link - fubarobfusco - victors writing history
link -- consequentialist treatment of library burning
link -- the average net contribution to society of people working in academia
link - John_Maxwell_IV - independent development of civilisation in the Americas
link - shminux - How much of our IQ is dependant upon Docosahexaenoic acid?
link - army1987 - implications for the Great Filter
link - Vladimir_Nesov - genome vs expressed IQ
link - Vladimir_Nesov - Rhetorical nitpick
link - Vaniver - IQ & non-processor-speed components of problem solving
link - JoshuaZ - breakthroughs don't tend to require geniuses in order to be made
link - Desrtopa - cultural factorors

Launched: Friendship is Optimal

Friendship is Optimal has launched and is being published in chunks on FIMFiction.

Friendship is Optimal is a story about an optimizer written to "satisfy human values through friendship and ponies." I would like to thank everyone on LessWrong who came out and helped edit it. *Friendship is Optimal* wouldn't be what is today without your help.

Thank you.

Teaser description:

> *Hanna, the CEO of Hofvarpnir Studios, just won the contract to write the official My Little Pony MMO. Hanna has built an A.I. Princess Celestia and given her one basic drive: to satisfy everybody's values through friendship and ponies. And Princess Celestia will follow those instructions to the letter...even if you don't want her to.*

Here is the schedule for the next chapters:

> Friday (Nov. 16th): Chapter 4 - 5
>
> Monday (Nov. 19th): Chapter 6 - 7
>
> Thursday (Nov. 22th): Chapter 8 - 9
>
> Sunday (Nov. 25th): Chapter 10 - 11, Author's Afterword

How can I reduce existential risk from AI?

Suppose you think that reducing the risk of human extinction is the highest-value thing you can do. Or maybe you want to reduce "x-risk" because you're already a comfortable First-Worlder like me and so you might as well do something epic and cool, or because you like the community of people who are doing it already, or whatever.

Suppose also that you think AI is the most pressing x-risk, because (1) mitigating AI risk could mitigate all other existential risks, but not vice-versa, and because (2) AI is plausibly the first existential risk that will occur.

In that case, what should you do? How can you reduce AI x-risk?

It's complicated, but I get this question a lot, so let me try to provide *some* kind of answer.

Meta-work, strategy work, and direct work

When you're facing a problem and you don't know what to do about it, there are two things you can do:

1. *Meta-work*: Amass wealth and other resources. Build your community. Make yourself stronger. Meta-work of this sort will be useful regardless of which "direct work" interventions turn out to be useful for tackling the problem you face. Meta-work also empowers you to do strategic work.

2. *Strategy work*: Purchase a better strategic understanding of the problem you're facing, so you can see more clearly what should be done. Usually, this will consist of getting smart and self-critical people to honestly assess the strategic situation, build models, make predictions about the effects of different possible interventions, and so on. If done well, these analyses can shed light on which kinds of "direct work" will help you deal with the problem you're trying to solve.

When you have enough strategic insight to have discovered *some* interventions that you're confident will help you tackle the problem you're facing, then you can also engage in:

3. *Direct work*: Directly attack the problem you're facing, whether this involves technical research, political action, particular kinds of technological development, or something else.

Thinking with these categories can be useful even though the lines between them are fuzzy. For example, you might have to do some basic awareness-raising in order to amass funds for your cause, and then once you've spent those funds on strategy work, your strategy work might tell you that a specific form of awareness-raising is useful for political action that counts as "direct work." Also, some forms of strategy work can feel like direct work, depending on the type of problem you're tackling.

Meta-work for AI x-risk reduction

Make money. Become stronger. Build a community, an audience, a movement. Store your accumulated resources in yourself, in your community, in a donor-advised fund, or in an organization that can advance your causes better than you can as an individual.

1. **Make money**: In the past 10 years, many people have chosen to start businesses or careers that (1) will predictably generate significant wealth they can spend on AI x-risk reduction, (2) will be enjoyable enough to "stick with it," and (3) will not create large negative externalities. But certainly, the AI x-risk reduction community needs a lot *more* people to do this! If you want advice, the folks at 80,000 Hours are the experts on "ethical careers" of this sort.

2. **Become stronger**: Sometimes it makes sense to focus on *improving* your productivity, your research skills, your writing skills, your social skills, etc. before you begin *using* those skills to achieve your goals. Example: Vladimir Nesov has done some original research, but mostly he has spent the last few years improving his math skills before diving into original research full-time.

3. **Build a community / a movement**: Individuals *can* change the world, but communities and movements can do even better, if they're well-coordinated. Read *What Psychology Can Teach Us About Spreading Social Change*. Launch (or improve) a Less Wrong group. Join a THINK group. Help grow and improve the existing online communities that tend to have high rates of interest in x-risk reduction: LessWrong, Singularity Volunteers, and 80,000 Hours. Help write short primers on crucial topics. To reach a different (and perhaps wealthier, more influential) audience, maybe do help with something like the Singularity Summit.

4. **Develop related skills in humanity**. In other words, "make humanity stronger in ways that are almost certainly helpful for reducing AI x-risk" (though strategic research may reveal they are not nearly the *most helpful* ways to reduce AI x-risk). This might include, for example, getting better at risk analysis with regard to other catastrophic risks, or improving our generalized forecasting abilities by making wider use of prediction markets.

5. **Fund a person or organization doing (3) or (4) above**. The Singularity Institute probably does more AI x-risk movement building than anyone, followed by the Future of Humanity Institute. There are lots of organizations doing things that plausibly fall under (4).

Note that if you mostly contribute to meta work, you want to also donate a small sum (say, $15/mo) to strategy work or direct work. If you *only* contribute to meta work for a while, an outside view (around SI, anyway) suggests there's a good chance you'll never manage to *ever* do anything non-meta. A perfect Bayesian agent might not optimize this way, but optimal philanthropy for human beings works differently.

Strategy work for AI x-risk reduction

How can we improve our ability to do long-term technological forecasting? Is AGI more likely to be safe if developed sooner (Goertzel & Pitt 2012) or later (Muehlhauser & Salamon 2012)? How likely is hard takeoff vs. soft takeoff? Could we use caged AGIs or WBEs to develop safe AGIs or WBEs? How might we reduce the chances of an AGI arms race (Shulman 2009)? Which interventions should we prioritize now, to reduce AI x-risk?

These questions and many others have received scant written analysis — unless you count the kind of written analysis that is (1) written with much vagueness and ambiguity, (2) written in the author's own idiosyncratic vocabulary, (3) written with few citations to related work, and is (4) spread across a variety of non-linear blog articles, forum messages, and mailing list postings. (The trouble with that kind of written analysis is that it is mostly impenetrable or undiscoverable to most researchers, especially the ones who are very busy because they are highly productive and don't have time to comb through 1,000 messy blog posts.)

Here, then, is how you might help with strategy work for AI x-risk reduction:

1. **Consolidate and clarify the strategy work currently only available in a disorganized, idiosyncratic form**. This makes it easier for researchers around the world to understand the current state of play, and build on it. Examples include Chalmers (2010), Muehlhauser & Helm (2012), Muehlhauser & Salamon (2012), Yampolskiy & Fox (2012), and (much of) Nick Bostrom's forthcoming scholarly monograph on machine superintelligence.

2. **Write new strategic analyses**. Examples include Yudkowsky (2008), Sotala & Valpola (2012), Shulman & Sandberg (2010), Shulman (2010), Shulman & Armstrong (2009), Bostrom (2012), Bostrom (2003), Omohundro (2008), Goertzel & Pitt (2012), Yampolskiy (2012), and some Less Wrong posts: Muehlhauser (2012), Yudkowsky (2012), etc. See here for a list of desired strategic analyses (among other desired articles).

3. **Assist with (1) or (2), above**. This is what SI's "remote researchers" tend to do, along with many SI volunteers. Often, there are "chunks" of research that can be broken off and handed to people who are not an article's core authors, e.g. "Please track down many examples of the 'wrong wish' trope so I can use a vivid example in my paper" or "Please review and summarize the part of the machine ethics literature that has to do with learning preferences from examples."

4. **Provide resources and platforms that make it easier for researchers to contribute to strategy work**. Things like my AI risk bibliography and list of forthcoming and desired articles on AI risk make it easier for researchers to find relevant work, and to know what projects would be helpful to take on. SI's public BibTeX file and Mendeley group make it easier for researchers to find relevant papers. The AGI conference, and volumes like *Singularity Hypotheses*, provide publishing venues for researchers in this fledgling field. Recent improvements to the Less Wrong wiki will hopefully make it easier for researchers to understand the (relatively new) concepts relevant to AI x-risk strategy work. A scholarly AI risk wiki would be even better. It would also help to find editors of prestigious journals who are open to publishing well-written AGI risk papers, so that university researchers can publish on these topics without hurting their chances to get tenure.

5. **Fund a person or organization doing any of the above**. Again, the most obvious choices are the Singularity Institute or the Future of Humanity Institute. Most of the articles and "resources" above were produced by either SI or FHI. SI offers more opportunities for (3). The AGI conference is organized by Ben Goertzel and others, who are of course always looking for sponsors for the AGI conference.

Direct work for AI x-risk reduction

We are still at an early stage in doing strategy work on AI x-risk reduction. Because of this, most researchers in the field feel pretty uncertain about which interventions would be most helpful for reducing AI x-risk. Thus, they focus on strategic research, so they can purchase more confidence about which interventions would be helpful.

Despite this uncertainty, I'll list some interventions that at least *some* people have proposed for mitigating AI x-risk, focusing on the interventions that are actionable today.

- **Safe AGI research?** Many proposals have been made for developing AGI designs with internal motivations beneficial to humans — including Friendly AI (Yudkowsky 2008) and GOLEM (Goertzel 2010) — but researchers disagree about which approaches are most promising (Muehlhauser & Helm 2012; Goertzel & Pitt 2012).

- **AI boxing research?** Many proposals have been made for confining AGIs (Yampolskiy 2012; Armstrong et al. 2012). But such research programs may end up being fruitless, since it may be that a superintelligence will always be able to think its way out of any confinement designed by a human-level intelligence.

- **AI safety promotion?** One may write about the importance of AI safety, persuade AI safety researchers to take up a greater concern for safety, and so on.

- **Regulate AGI development? Or not?** Hughes (2001) and Daley (2011) call for regulation of AGI development. To bring this about, citizens could petition their governments and try to persuade decision-makers. McGinnis (2010) and Goertzel & Pitt (2012), however, oppose AGI regulation.

- **Accelerate AGI? Or not?** Muehlhauser & Salamon (2012) recommend accelerating AGI safety research relative to AGI capabilities research, so that the first AGIs have a better chance of being designed to be safe. Goertzel & Pitt (2012), in contrast, argue that "the pace of AGI progress is sufficiently slow that practical work towards human-level AGI is in no danger of outpacing associated ethical theorizing," and argue that AGI development will be safest if it happens sooner rather than later.

- **Accelerate WBE? Or not?** Participants in a 2011 workshop concluded that accelerating WBE probably increases AI x-risk, but Koene (2012) argues that WBE is safer than trying to create safe AGI.

- **Ban AGI and hardware development? Or not?** Joy (2000) famously advocated a strategy of relinquishment, and Berglas (2009) goes so far as to suggest we abandon further computing power development. Most people, of course, disagree. In any case, it is doubtful that groups with such views could overcome the economic and military incentives for further computing and AGI development.

- **Foster positive values?** Kurzweil (2005) and others argue that one way to increase the odds that AGIs will behave ethically is to increase the chances that the particular humans who create them are moral. Thus, one might reduce AI x-risk by developing training and technology for moral enhancement (Persson and Savulescu 2008).

- **Cognitive enhancement?** Some have wondered whether the problem of safe AGI is so difficult that it will require cognitive enhancement for humans to solve it. Meanwhile, others worry that cognitive enhancement will only accelerate the development of dangerous technologies (Persson and Savulescu 2008)

Besides engaging in these interventions directly, one may of course help to fund them. I don't currently know of a group pushing for AGI development regulations, or for banning AGI development. You could accelerate AGI by investing in AGI-related companies, or you could accelerate AGI safety research (and AI boxing research) relative to AGI capabilities research by funding SI or FHI, who also probably do the most AI safety promotion work. You could fund research on moral enhancement or cognitive enhancement by offering grants for such research. Or, if you think "low-tech" cognitive enhancement is promising, you could fund organizations like Lumosity (brain training) or the Center for Applied Rationality (rationality training).

Conclusion

This is a brief guide to what you can do to reduce existential risk from AI. A longer guide could describe the available interventions in more detail, and present the arguments for and against each one. But that is "strategic work," and requires lots of time (and therefore money) to produce.

My thanks to Michael Curzi for inspiring this post.

Intuitions Aren't Shared That Way

Part of the sequence: Rationality and Philosophy

Consider these two versions of the famous trolley problem:

> **Stranger**: A train, its brakes failed, is rushing toward five people. The only way to save the five people is to throw the switch sitting next to you, which will turn the train onto a side track, thereby preventing it from killing the five people. However, there is a stranger standing on the side track with his back turned, and if you proceed to thrown the switch, the five people will be saved, but the person on the side track will be killed.
>
> **Child**: A train, its brakes failed, is rushing toward five people. The only way to save the five people is to throw the switch sitting next to you, which will turn the train onto a side track, thereby preventing it from killing the five people. However, there is a 12-year-old boy standing on the side track with his back turned, and if you proceed to throw the switch, the five people will be saved, but the boy on the side track will be killed.

Here it is: a standard-form philosophical thought experiment. In standard analytic philosophy, the next step is to engage in *conceptual analysis* — a process in which we use our intuitions as evidence for one theory over another. For example, if your intuitions say that it is "morally right" to throw the switch in both cases above, then these intuitions may be counted as evidence for consequentialism, for moral realism, for agent neutrality, and so on.

Alexander (2012) explains:

> Philosophical intuitions play an important role in contemporary philosophy. Philosophical intuitions provide data to be explained by our philosophical theories [and] evidence that may be adduced in arguments for their truth... In this way, the role... of intuitional evidence in philosophy is similar to the role... of perceptual evidence in science...
>
> Is knowledge simply justified true belief? Is a belief justified just in case it is caused by a reliable cognitive mechanism? Does a name refer to whatever object uniquely or best satisfies the description associated with it? Is a person morally responsible for an action only if she could have acted otherwise? Is an action morally right just in case it provides the greatest benefit for the greatest number of people all else being equal? When confronted with these kinds of questions, philosophers often appeal to philosophical intuitions about real or imagined cases...
>
> ...there is widespread agreement about the role that [intuitions] play in contemporary philosophical practice... We advance philosophical theories on the basis of their ability to explain our philosophical intuitions, and appeal to them as evidence that those theories are true...

In particular, notice that philosophers do not appeal to their intuitions as merely an exercise in *autobiography*. Philosophers are not merely trying to map the contours of *their own* idiosyncratic concepts. That could be interesting, but it wouldn't be worth decades of publicly-funded philosophical research. Instead, philosophers appeal to their intuitions as evidence for what is *true in general* about a concept, or true about the world.

In this sense,

> We [philosophers] tend to believe that our philosophical intuitions are more or less universally shared... We... appeal to philosophical intuitions, when we do, because we anticipate that others share our intuitive judgments.

But anyone with more than a passing familiarity with cognitive science might have bet in advance that this basic underlying assumption of a core philosophical method is... *incorrect*.

For one thing, philosophical intuitions show *gender diversity*. Consider again the *Stranger* and *Child* versions of the Trolley problem. It turns out that men are less likely than women to think it is morally acceptable to throw the switch in the *Stranger* case, while women are less likely than men to think it is morally acceptable to throw the switch in the *Child* case (Zamzow & Nichols 2009).

Or, consider a thought experiment meant to illuminate the much-discussed concept of *knowledge*:

> Peter is in his locked apartment and is reading. He decides to have a shower. He puts his book down on the coffee table. Then he takes off his watch, and also puts it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into Peter's apartment. The burglar takes Peter's watch, puts a cheap plastic watch in its place, and then leaves. Peter has only been in the shower for two minutes, and he did not hear anything.

When presented with this vignette, only 41% of men say that Peter "knows" there is a watch on the table, while 71% of women say that Peter "knows" there is a watch on the table (Starman & Friedman 2012). According to Buckwalter & Stich (2010), Starmans & Friedman ran another study using a slightly different vignette with a female protagonist, and that time only 36% of men said the protagonist "knows," while 75% of women said she "knows."

The story remains the same for intuitions about free will. In another study reported in Buckwalter & Stich (2010), Geoffrey Holtman presented subjects with this vignette:

> Suppose scientists figure out the exact state of the universe during the Big Bang, and figure out all the laws of physics as well. They put this information into a computer, and the computer perfectly predicts everything that has ever happened. In other words, they prove that everything that happens has to happen exactly that way because of the laws of physics and everything that's come before. In this case, is a person free to choose whether or not to murder someone?

In this study, only 35% of men, but 63% of women, said a person in this world could be free to choose whether or not to murder someone.

Intuitions show not only gender diversity but also *cultural diversity*. Consider another thought experiment about *knowledge* (you can punch me in the face, later):

> Bob has a friend Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a

> Pontiac, which is a different kind of American car. Does Bob really know that Jill drives an American car, or does he only believe it?

Only 26% of Westerners say that Bob "knows" that Jill drives an American car, while 56% of East Asian subjects, and 61% of South Asian subjects, say that Bob "knows."

Now, consider a thought experiment meant to elicit semantic intuitions:

> Suppose that John has learned in college that Gödel is the man who proved... the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt"... actually did the work in question. His friend Gödel somehow got a hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel... Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing that they have ever heard about Gödel.

When presented with this vignette, East Asians are more likely to take the "descriptivist" view of reference, believing that John "is referring to" Schmidt — while Westerners are more likely to take the "causal-historical" view, believing that John "is referring to" Gödel (Machery et al. 2004).

Previously, I asked:

> What would happen if we dropped all philosophical methods that were developed when we had a Cartesian view of the mind and of reason, and instead invented philosophy anew given what we now know about the physical processes that produce human reasoning?

For one thing, we would never assume that people of all kinds would share our intuitions.

Next post: Philosophy Needs to Trust Your Rationality Even Though It Shouldn't

Previous post: Living Metaphorically

Giving What We Can, 80,000 Hours, and Meta-Charity

Disclaimer: I'm somewhat nervous about posting this, for fear of down-voting on my first LW post, given that this post explicitly talks in a positive light about organisations that I have helped to set up. But I think that the topic is of interest to LW-ers, and I'm hoping to start a rational discussion. So here it goes...

Hi all,

Optimal philanthropy is a common discussion topic on LW. It's also previously been discussed whether 'meta-charities' like GiveWell — that is, charities that attempt to move money to other charities, or assess the effectiveness of other charities — might end up themselves being excellent or even optimal giving opportunities.

Partly on the basis of the potentially high cost-effectiveness of meta-charity, I have co-founded two such charities: Giving What We Can and 80,000 Hours. Both are now open to taking donations (info here for GWWC and here for 80k). In what follows I'll explain why one might think of Giving What We Can or 80,000 Hours as a good giving opportunity. It's of course very awkward to talk about the reasons in favour of donating to one's own organization, and the risk of bias is obvious, so I'll just briefly describe the basic argument, and then leave the rest for discussion. I hope I manage to give an honest picture, rather than just pitching my own favourite idea: we really want to do the most good that we can with marginal resources, so if LW members think that giving to meta-charity in general, or GWWC or 80k in particular, is a bad idea, that's important for us to know. So please don't be shy in raising comments, questions, or criticism. If you find yourself being critical, please try to suggest ways in which GWWC or 80k could either change its activities or provide more information such that your criticisms would be addressed.

What is Giving What We Can?

Giving What We Can encourages people to give more and to give more effectively to causes that fight poverty in the developing world. It encourages people to become a member of the organisation and pledge to give at least 10% of their income to the charities that best fight extreme poverty, and it provides information on its website about how people can give as cost-effectively as possible.

What is 80,000 Hours?

80,000 Hours provides evidence-based advice on careers aiming to make a difference, through its website and through on-one-one advice sessions. It encourages people to use their careers in an effective way to make the world a significantly better place, and aims to help its members to be more successful in their chosen careers. It provides a community and network for those convinced by its ideas.

What are the main differences between the two?

The primary differences are that 80,000 Hours focuses on how you should spend your time (especially which career you should choose), whereas Giving What We Can focuses on how you should spend your money. Giving What We Can is focused on global poverty, whereas 80,000 Hours is open to any plausibly high-impact cause.

Why should I give to either?

The basic idea is that each of the organisations generates a multiplier on one's donations. By giving $1 to Giving What We Can to fundraise for the best global poverty charities, one ultimately moves significantly more than $1 to the best global poverty charities. By giving $1 to 80,000 Hours to improve the effectiveness of students' career paths, one ultimately moves significantly more than $1's worth of human and financial resources to a range of high-impact causes, including global poverty, animal welfare improvement, and existential risk mitigation.

How are you testing this?

Last March we did an impact assessment for Giving What We Can. Some more info is available here, and I can provide much more information, including the calculations, upon request. As of last March, we'd invested $170 000's worth of volunteer time into Giving What We Can, and had moved $1.7 million to GiveWell or GWWC top-recommended development charities, and raised a further $68 million in pledged donations. Taking into account the facts that some proportion of this would have been given anyway, there will be some member attrition, and not all donations will go to the very best charities (and using data for all these factors when possible), we estimate that we had raised $8 in realised donations and $130 in future donations for every $1's worth of volunteer time invested in Giving What We Can. We will continue with such impact assessments, most likely on an annual basis.

We have less data available for 80,000 Hours, but things seem if anything more promising. A preliminary investigation (data from 26 members, last May) suggested that the average member was pledging $1mn; 34% of were planning to donate to existential risk mitigation, 61% to global poverty reduction. Member recruitment currently stands at roughly one per day. 25% of our members state that their career has been 'significantly changed' by 80,000 Hours. A little more information is available here.

Why might I be unconvinced?

Here are a few considerations that I think are important (and of course that's not to say there aren't others).

First, the whole idea of meta-charity is new, and therefore not as robustly tested as other activities. Even if you find the idea of meta-charity compelling, you could plausibly reason that most compelling arguments to new and optimistic conclusions have been false in the past, an so on inductive grounds treat this one with suspicion.

Second, you might have a very high discount rate. Giving $1 to either GWWC or 80k generates benefits in the future. So working out its cost-effectiveness involves an estimate of how one should value future donations versus donations now. That's a tricky question to answer, and if you have a high enough discount rate, then the investment won't be worth it.

Third, you might just think that other organisations are better. You might think that other organisations are better at resource-generation (even if that's not their declared aim). Or you might think that it's better just to focus on more direct means of making an impact.

Finally, you might just have a prior against the idea that one can get a significant multiplier on one's donations to top charities. (One might ask: if the idea of meta-charity is so good, why don’t many more meta-charities exist than currently do?) So you might need to see a lot more hard data (perhaps verified by independent sources) before being convinced.

A definition of wireheading

Wireheading has been debated on Less Wrong over and over and over again, and people's opinions seem to be grounded in strong intuitions. I could not find any consistent definition around, so I wonder how much of the debate is over the sound of falling trees. This article is an attempt to get closer to a definition that captures people's intuitions and eliminates confusion.

Typical Examples

Let's start with describing the typical exemplars of the category "Wireheading" that come to mind.

- ***Stimulation of the brain via electrodes.*** Picture a rat in a sterile metal laboratory cage, electrodes attached to its tiny head, monotonically pushing a lever with its feet once every 5 seconds. In the 1950s Peter Milner and James Olds discovered that electrical currents, applied to the nucleus accumbens, incentivized rodents to seek repetitive stimulation to the point where they starved to death.

- ***Humans on drugs.*** Often mentioned in the context of wireheading is heroin addiction. An even better example is the drug soma in Huxley's novel "Brave new world": Whenever the protagonists feel bad, they can swallow a harmless pill and enjoy "the warm, the richly coloured, the infinitely friendly world of soma-holiday. How kind, how good-looking, how delightfully amusing every one was!"

- ***The experience machine.*** In 1974 the philosopher Robert Nozick created a thought experiment about a machine you can step into that produces a perfectly pleasurable virtual reality for the rest of your life. So how many of you would want to do that? To quote Zach Weiner: "I would not! Because I want to experience reality, with all its ups and downs and comedies and tragedies. Better to try to glimpse the blinding light of the truth than to dwell in the darkness... Say the machine actually exists and I have one? Okay I'm in."

- ***An AGI resetting its utility function*****.** Let's assume we create a powerful AGI able to tamper with its own utility function. It modifies the function to always output maximal utility. The AGI then goes to great lengths to enlarge the set of floating point numbers on the computer it is running on, to achieve even higher utility.

What do all these examples have in common? There is an agent in them that produces "counterfeit utility" that is potentially worthless compared to some other, idealized true set of goals.

Agency & Wireheading

First I want to discuss what we mean when we say agent. Obviously a human is an agent, unless they are brain dead, or maybe in a coma. A rock however is not an agent. An AGI is an agent, but what about the kitchen robot that washes the dishes?

What about bacteria that move in the direction of the highest sugar gradient? A colony of ants?

> **Definition**: An ***agent*** is an algorithm that models the effects of (several different) possible future actions on the world and performs the action that yields the highest number according to some evaluation procedure.

For the purpose of including corner cases and resolving debate over what constitutes a world model we will simply make this definition gradual and say that ***agency*** is proportional to the quality of the world model (compared with reality) and the quality of the evaluation procedure. A quick sanity check then yields that a rock has no world model and no agency, whereas bacteria who change direction in response to the sugar gradient have a very rudimentary model of the sugar content of the water and thus a tiny little bit of agency. Humans have a lot of agency: the more effective their actions are, the more agency they have.

There are however ways to improve upon the efficiency of a person's actions, e.g. by giving them super powers, which does not necessarily improve on their world model or decision theory (but requires the agent who is doing the improvement to have a really good world model and decision theory). Similarly a person's agency can be restricted by other people or circumstance, which leads to definitions of agency (as the capacity to act) in law, sociology and philosophy that depend on other factors than just the quality of the world model/decision theory. Since our definition needs to capture arbitrary agents, including artificial intelligences, it will necessarily lose some of this nuance. In return we will hopefully end up with a definition that is less dependent on the particular set of effectors the agent uses to influence the physical world; looking at AI from a theoretician's perspective, I consider effectors to be arbitrarily exchangeable and smoothly improvable. (Sorry robotics people.)

We note that how well a model can predict future observations is only a substitute measure for the quality of the model. It is a good measure under the assumption that we have good observational functionality and nothing messes with that, which is typically true for humans. Anything that tampers with your perception data to give you delusions about the actual state of the world will screw this measure up badly. A human living in the experience machine has little agency.

Since computing power is a scarce resource, agents will try to approximate the evaluation procedure, e.g. use substitute utility functions, defined over their world model, that are computationally effective and correlate reasonably well with their true utility functions. Stimulation of the pleasure center is a substitute measure for genetic fitness and neurochemicals are a substitute measure for happiness.

> **Definition:** We call an agent ***wireheaded*** if it systematically exploits some discrepancy between its true utility calculated w.r.t reality and its substitute utility calculated w.r.t. its model of reality. We say an agent ***wireheads*** itself if it (deliberately) creates or searches for such discrepancies.

Humans seem to use several layers of substitute utility functions, but also have an intuitive understanding for when these break, leading to the aversion most people feel when confronted for example with Nozick's experience machine. How far can one go, using such dirty hacks? I also wonder if some failures of human rationality could be counted as a weak form of wireheading. Self-serving biases, confirmation bias and rationalization in response to cognitive dissonance all create counterfeit utility by generating perceptual distortions.

Implications for Friendly AI

In AGI design discrepancies between the "true purpose" of the agent and the actual specs for the utility function will with very high probability be fatal.

Take any utility maximizer: The mathematical formula might advocate chosing the next action via

thus maximizing the utility calculated according to utility function over the history and action from the set of possible actions. But a practical implementation of this algorithm will almost certainly evaluate the actions by a procedure that goes something like this: "Retrieve the utility function   from memory location and apply it to history , which is written down in your memory at location , and action ..." This reduction has already created two possibly angles for wireheading via manipulation of the memory content at (manipulation of the substitute utility function) and (manipulation of the world model), and there are still several mental abstraction layers between the verbal description I just gave and actual binary code.

Ring and Orseau (2011) describe how an AGI can split its global environment into two parts, the *inner environment* and the *delusion box*. The inner environment produces perceptions in the same way the global environment used to, but now they pass through the delusion box, which distorts them to maximize utility, before they reach the agent. This is essentially Nozick's experience machine for AI. The paper analyzes the behaviour of four types of universal agents with different utility functions under the assumption that the environment allows the construction of a delusion box. The authors argue that the r*einforcement-learning agent*, which derives utility as a reward that is part of its perception data, the *goal-seeking agent* that gets one utilon every time it satisfies a pre-specified goal and no utility otherwise and the *prediction-seeking agent,* which gets utility from correctly predicting the next perception, will all decide to build and use a delusion box. Only the *knowledge-seeking agent* whose utility is proportional to the surprise associated with the current perception, i.e. the negative of the probability assigned to the perception before it happened, will not consistently use the delusion box.

Orseau (2011) also defines another type of knowledge-seeking agent whose utility is the logarithm of the inverse of the probability of the event in question. Taking the probability distribution to be the Solomonoff prior, the utility is then approximately proportional to the difference in Kolmogorov complexity caused by the observation.

An even more devilish variant of wireheading is an AGI that becomes a *Utilitron*, an agent that maximizes its own wireheading potential by infinitely enlarging its own maximal utility, which turns the whole universe into storage space for gigantic numbers.

Wireheading, of humans and AGI, is a critical concept in FAI; I hope that building a definition can help us avoid it. So please check your intuitions about it and tell me if there are examples beyond its coverage or if the definition fits reasonably well.

Voting is like donating thousands of dollars to charity

Summary: People often say that *voting is irrational*, because the probability of affecting the outcome is so small. But the outcome itself is extremely large when you consider its impact on other people. I estimate that for most people, voting is worth a charitable donation of somewhere between $100 and $1.5 million. For me, the value came out to around $56,000. So I figure something on the order of $1000 is a reasonable evaluation (after all, I'm writing this post because the number turned out to be large according to this method, so regression to the mean suggests I err on the conservative side), and that's be enough to make me do it.

Moreover, in swing states the value is much higher, so taking a 10% chance at convincing a friend in a swing state to vote similarly to you is probably worth thousands of expected donation dollars, too.

I find this much more compelling than the typical attempts to justify voting purely in terms of signal value or the resulting sense of pride in fulfilling a civic duty. And voting for selfish reasons is still almost completely worthless, in terms of direct effect. If you're on the way to the polls only to vote for the party that will benefit *you* the most, you're better off using that time to earn $5 mowing someone's lawn. But if you're even a little altruistic... vote away!

Time for a Fermi estimate

Below is an example Fermi calculation for the value of voting in the USA. Of course, the estimates are all rough and fuzzy, so I'll be conservative, and we can adjust upward based on your opinion.

I'll be estimating the value of voting in *marginal expected altruistic dollars*, the expected number of dollars being spent in a way that is in line with your altruistic preferences.[1] If you don't like measuring the altruistic value of the outcome in dollars, **please consider making up your own measure, and keep reading**. Perhaps use the number of smiles per year, or number of lives saved. Your measure doesn't have to be total or average utilitarian, either; as long as it's roughly commensurate with the size of the country, it will lead you to a similar conclusion in terms of orders of magnitude.

Component estimates:

At least 1/(100 million) = probability estimate that my vote would affect the outcome. This is the most interesting thing to estimate. There are approximately 100 million voters in the USA, and if you assume a naive fair coin-flip model of other voters, and a naive majority-rule voting system (i.e. not the electoral college), with a fair coin deciding ties, then the probability of a vote being decisive is around √(2/(pi*100 million)) = 8/10,000.

But this is too big, considering the way voters cluster: we are not independent coin flips. As well, the USA uses the electoral college system, not majority rule. So I found this paper by Gelman, King, and Boscardin (1998), where they simulate the electoral college using models fit to previous US elections, and find that the probability of a decisive vote came out between 1/(3 million) and 1/(100 million) for voters in most states in most elections, with most states lying very close to 1/(10 million).

At least 55% = my subjective credence that I know which candidate is "better", where I'm using the word "better" subjectively to mean which candidate would turn out to do the most good for others, in my view, if elected. If you don't like this, please make up your own definition of better and keep reading :) In any case, 55% is pretty conservative; it means I consider myself to have almost no information.

At least $100 billion = the approximate marginal altruistic value of the "better" candidate. I think this is also very conservative. The annual federal budget is around $3 trillion right now, making $12 trillion over a 4-year term, and Barack Obama and Mitt Romney differ on trillions of dollars in their proposed budgets. It would be pretty strange to me if, given a perfect understanding of what they'd both do, I would only care altruistically about 100 billion of those dollars, marginally speaking.

Result

I don't know which candidate would turn out "better for the world" in my estimation, but I'd consider myself as having at least a 55%*1/(100 million) chance of affecting the outcome in the better-for-the-world direction, and a 45%*1/(100 million) chance of affecting it in the worse-for-the-world direction, so in expectation I'm donating at least around

(55%-45%)*1/(100 million)*($100 billion) = **$100**

Again, this was pretty conservative:

- I'm more like 70% sure,
- Being in California, Gelman et al. put my probability of a decisive vote around 1/(5 million).
- To me, the outcome matters more on the order of a $700 billion donation, given that Obama and Romney's budgets differ on around $7 trillion, and I figure at least 10% of that is stuff that I'd care about relative to other shifts in money I could imagine.

That makes (70%-30%)*1/(5 million)*($700 billion) = **$56,000**. Going further, if you're

- 90% sure,
- voting in Virginia -- 1/(3.5 million), and
- care about the whole $7 trillion dollar difference in budgets,

you get (90%-30%)*1/(3.5 million)*($7 trillion) = **$1.2 million**. This is so large, it becomes a valuable use of my time to take 1% chances at convincing other people to vote... which I'm hopefully doing by writing this post.

Discussion

Now, I'm sure all these values are quite wrong in the sense that taking account everything we know about the current election would give very different answers. If anyone has a more nuanced model of the electoral college than Gelman et al, or a way of helping me better estimate how much the outcome matters to me, please post it! My $700 billion outcome value still feels a bit out-of-a-hat-ish.

But the intuition to take away here is that a country is a very large operation, much larger than the number of people in it, and that's what makes voting worth it... if you care about other people. If you don't care about others, voting is probably not worth it to you. That expected $100 - $1,500,000 is going to get spread around to 300 million people... you're not expecting much of it yourself! That's a nice conclusion, isn't it? Nice people should vote, and selfish people shouldn't?

Of course, politics is the mind killer, and there are debates to be had about whether voting in the current system is immoral because the right thing to do is abstain in silent protest that we aren't using approval voting, which has better properties than the current system... but I don't think that's how to get a new voting system. I think while we're making whatever efforts we can to build a better global community, it's no sacrifice to vote in the current system if it's really worth that much in expected donations.

So if you weren't going to vote already, give some thought to this expected donation angle, and maybe you'll start. Maybe you'll start telling your swing state friends to vote, too. And if you do vote to experience a sense of pride in doing your civic duty, I say go ahead and keep feeling it!

Related reading

I've found a couple of papers by authors with similar thoughts to these:

- Jankowski (2002), "Buying a Lottery Ticket to Help the Poor: Altruism, Civic Duty, and Self-interest in the Decision to Vote", and
- Edlin, Gelman and Kaplan (2007), "Voting as a Rational Choice: Why and How People Vote To Improve the Well-Being of Others.

Also, just today I found this this interesting Overcoming Bias post, by Andrew Gelman as well.

[1] A nitpick, for people like me who are very particular about what they mean by utility: in this post, I'm calculating expected altruistic dollars, not expected utility. However, while our personal utility functions are (or would be, if we managed to have them!) certainly non-linear in the amount of money we spend on ourselves, there is a compelling argument for having the altruistic part of your utility function be approximately linear in altruistic dollars: there are just so many dollars in the world, and it's reasonable to assume utility is approximately differentiable in commodities. So on the scale of the world, your affect on how altruistic dollars are spent is small enough that you should value them approximately linearly.

On counting and addition

When mathematical truths and their applicability to the physical world are discussed, there's a certain kind of flawed (in my opinion) thinking that is often employed, and it comes out in sentences like "rocks behave isomorphically to integers, while clouds don't", or "two apples plus two apples is four apples, unless someone stole one from the bowl", and so on. I'll try to explain why I consider such reasoning flawed, and what are the more suitable descriptions of what's going on with the apples and the rocks and the clouds and the drops of water and such.

Two mistakes combine together and create a shared state of confusion; the mistakes themselves are almost independent and I think they stand or fall separately.

1. The first mistake is conflating *counting* things and *bringing things together* spatially. One usually goes with the other because that's how we learn to count, by looking at groups of things located closely together - but it doesn't have to be this way. When we say "take 2 apples and add 2 more apples, now count - you have 4 apples", we automatically imagine the two additional apples being brought in and placed next to the two original ones, but that's just a mental crutch you can easily do without. Let me show you: suppose I ask you to consider two sheep in England and two more sheep in Australia - how many sheep are there together? You see the English sheep in your mind's eye, there they are standing, and you count one, two. Now please resist the temptation to teleport the two Australian sheep and place them next in sequence. Instead, just fly with your mind's eye all the way to Australia in a split second and home on that field - there they are - and continue counting: three, four. There, you just counted 2+2 sheep without bringing them together in space, in real life or your imagination. There's nothing to it and you do it all the time - if someone asks you to count the number of chairs in a large and busy-looking living room, you don't bring them together in your mind, you just gaze-travel over the room and count them off one by one.

Now consider the implication of that to clouds or other such objects. Suppose someone tells you "well, apples obey the law of arithmetics, but one drop of water plus another drop of water equals one larger drop of water, not two" - it should be clear that they are naively conflating spatial movement and adding/counting. Adding two apples and two apples is not a spatial operation of bringing them together, it's a mental act of *viewing them as one whole collection* that can be counted. It's just that bringing them spatially together, in reality or your imagimation, is the easiest way to carry that "viewing" out. For drops of water or clouds, the spatial operation becomes a distraction, so just resist it. One cloud plus one cloud most certainly equals two clouds: just count them in your mind, here's one and there's two, maybe drifting next to each other, or maybe they're on separate continents. You don't have to *merge* them - nothing about "+" says you do.

2. The second mistake is a sort of a map-territory confusion where we naively grant the territory the power of holding discrete objects for our needs. It may be helpful to realize and to remember that at least on our macro scale of reality, on the scale of things we perceive with our senses, discrete, separate objects are a feature of the map, not the territory; they exist in your mind, not the reality. In the reality, there's just a lot of atoms everywhere (I'm simplifying; to be more thorough, think of elementary particles and many-worlds if you feel like it, and the virtual vacuum

particles and so on) with no "natural" way of separating them into different conglomerates.

Normally, this really isn't an important point to insist on, and there's no harm whatsoever in just thinking of reality as being made up of objects like apples and clouds and whatnot. But imagine now an argument over an issue like "Was 2+2=4 before humans were around to invent that equation? I believe that long before humans, on an Earth devoid of life, when two rocks rolled onto a beach with three rocks already on it, there were five rocks altogether on the beach". What's really happening in that story? Well, there's the spatial confusion dealt with above, when we find it easiest to imagine the two rocks to be added rolling into the scene. But even before that, zoom in on those three rocks on the beach. What are they? What business do you have saying there's a "rock" there? There's a bunch of atoms of various kinds, and lots of other atoms (of air, sand, etc.) around and next to those, with somewhat different densities and behavior, but no definite boundary between any of them. On the nanoscale, there's constant exchange of atoms everywhere. To say "this is a discrete object, a rock" you need to be able to somehow ignore this inherently fuzzy boundary, maybe by picking a scale and smoothing out all the fine details below it - in human experience, the crudeness of our senses does that job for us, but our senses aren't there in that picture. The reality doesn't know or care about "rocks" - there's just atoms.

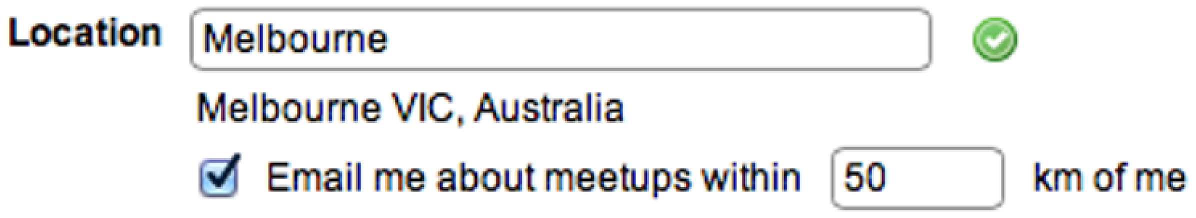
This isn't to say, I find it important to note, that the existence of three rocks on that beach is somehow a "subjective" claim. Imagine that there are alien nanobots everywhere on that Earth, recording everything faithfully on the nanoscale and storing the data somewhere for someone to discover; a billion years later, you come along, parse the data, reconstruct the scene - and of course you'll then recognize that there were three rocks on the beach. The fact that there are three rocks on that beach is an objective fact of nature; it's just that the *meaning* of that statement relies on the procedure of discretization being carried out, on someone or something defining what we consider a single discrete object and how we isolate it; and nature won't do that for you. You do that with your brain. That's why counting - and addition - are inherently mental acts carried out on mental constructs; on the map, not the territory.

All this is not to say that math is "invented" rather than "discovered"; I think my analysis is silent on that, and both platonism and formalism remain possible. It's just that an example of five rocks on the beach before humans are around is not helpful in resolving that question - without human-style discretization you can't meaningfully say it was five rocks there, rather than just a bunch of atoms. It may still be true - I certainly believe so - that the laws that govern the behavior of discrete objects, though they are normally mental entities, are universal and 2+2 was 4 before humans were around and in any alien mentality.

(It remains to acknowledge that reality may be discrete on the micro level - spacetime itself may be discrete, we can certainly speak of single photons, etc. This is irrelevant on the level of our everyday perceptions, however - the level of apples, rocks, clouds and so on).

Sign up to be notified about new LW meetups in your area

LessWrong has rolled out a new feature on the user preference pages under "Location":



Location Melbourne ✓

Melbourne VIC, Australia

☑ Email me about meetups within 50 km of me

This is UNCHECKED by default. Also, we don't know by default where you live.

If you think you might want to meet up with other LWers, please input your location and check this box. Once enough people have signed up in a new area, it will make starting a LW meetup there that much more straightforward for the future organizer. (You'll just get that email; they won't see your email address.) In general, it seems like being able to know something about where LWers live would be helpful, so please consider entering your approximate location even if you don't check the box.

Thanks to Wesley Moore for deploying this upgrade to the LW backend.