Best of LessWrong: December 2012

Best of LessWrong: December 2012

What science needs

Science does not need more scientists. It doesn't even need you, brilliant as you are. We already have many times more brilliant scientists than we can fund. Science could use a better understanding of the scientific method, but improving how individuals do science would not address most of the problems I've seen.

The big problems facing science are organizational problems. We don't know how to identify important areas of study, or people who can do good science, or good and important results. We don't know how to run a project in a way that makes correct results likely. Improving the quality of each person on the project is not the answer. The problem is the system. We have organizations and systems that take groups of brilliant scientists, and motivate them to produce garbage.

I haven't got it all figured out, but here are some of the most-important problems in science. I'd like to turn this into a front-page post eventually, but now I'm going to post it to discussion, and ask you to add new important problems in the comments.

Egos

A lot of LWers think they want to advance scientific understanding. But I've learned after years in the field that what most scientists want even more is prove how smart they are.

I couldn't tell you how many times I've seen a great idea killed because the project leader or someone else with veto power didn't want someone else's idea or someone else's area of expertise to appear important. I've been "let go" from two jobs because I refused when my bosses flat-out told me to stop proposing solutions for the important problems, because that was their territory.

I don't mean that you should try to stop people from acting that way. People act that way. I mean you should admit that people act that way, and structure contracts, projects, and rewards so that these petty ego-boosts aren't the biggest rewards people can hope to get.

Too many "no"-men

The more people your project has who can say "no", the worse the results will be. This is one reason why Hollywood feature films are stupid, why start-ups do good work, and why scientific projects are so often a waste of money. Good ideas are inherently unpopular. Most of the projects that I've worked on have been crippled because every good idea ran into someone with veto power who didn't want to do things differently, or didn't want somebody else to get credit for solving the problem. See "Egos".

Saying "no" to bad projects is important, but once the project is underway, there is a bias to say "no" more than "yes", even after adjusting for the number of times you can say "yes" in total. Requiring consensus is especially pernicious. You can't get good results when everbody on the project has to say "yes" to new ideas.

Jurisdiction arguments

Team members often disagree about whose expertise particular decisions fall under. Most people see how their expertise applies to a problem more easily than they can see how someone else's expertise applies to a problem. What usually happens is that territorial claims are honored from the top of the org chart on down, and by seniority. For example, I worked for a computer game company where the founder hired a scriptwriter, then came up with his own story ideas and told the scriptwriter to implement them. The implementation had no text; the scriptwriter took the story ideas and produced descriptions of scenes acted out with body language. The animators thought that body motion fell completely within their jurisdiction, so they felt free to rework whatever they saw differently. The scriptwriter had very little chance for creative input, no control over anything, and very little job satisfaction.

This is a common problem for computer scientists and mathematicians. Computer scientists and mathematicians see themselves as people who understand how to most-effectively take a set of data, and arrive at the desired results. This includes figuring out what data to look at, and in the best case, means being involved in the proposal writing to look at possible problems to address, and determine which problems are soluble and which ones are not based on information theory. This never happens. People in other specialties see computer scientists as a kind of lab technician to bring on after they've figured out what problem to address, and what data and general algorithm to use. They see statisticians as people to consult when the project is done and they're writing up the results. They aren't even aware that these other disciplines can do more than that.

A classic example is the Human Genome Project. Some people you never hear about, including my current boss, came up with algorithms to take whole-genome shotgun data and assemble it. Craig Venter went to the leaders of the Human Genome Project and explained to them that, using this approach, they could finish the project at a fraction of the cost. Anybody with a little mathematical expertise could look at the numbers and figure out on the back of a napkin that, yes, this could work. But all the decision-makers on the HGP were biologists. I presume that they didn't understand the math, and didn't believe that mathematicians could have useful insights into biological problems. So they declared it impossible—not difficult, but theoretically impossible—and plowed ahead, while Craig split off to use the shotgun approach. Billions of taxpayer dollars were wasted because a few people in leadership positions could not recognize that a problem in biology had a mathematical aspect.

Muzzling the oxen

"Thou shalt not muzzle the ox when he treadeth out the corn." — Deuteronomy 25:4

I believe that a large number of the problems with scientific research are tolerated only because nothing is at stake financially. Government agencies have tried very hard to ensure that people do work for their contracts. You have to say in the proposal what you're going to do, and itemize all your costs, and do what you said you would do, and write reports once a month or once a quarter showing that you're doing what you said you would do. This results in unfortunate obvious stupidities. We can spend $30,000 to have an employee write a piece of software that we could have bought for $500, or to solve a problem that a consultant could have solved for $500, but we can't buy the software or hire the consultant because they aren't listed in the contract and the employee is.

But the bigger problem is that the strict financial structure of scientific research makes it illegal to motivate scientists by giving them a percentage of resulting profits. You simply can't write up a budget proposal that way. So managers and team members indulge their prejudices and fantasies because the little bit of self-esteem boost they get from clinging to their favorite ideas is worth more to them than the extra money they would earn (zero) if the project produced better results. Examples of petty prejudices that I've seen people wreck good work to preserve: top-down over bottom-up design, emacs over vim (I was in a shop once where the founders forbade people from using vim, which had an astonishingly destructive effect on morale), rule-based over statistical grammars, symbolic logic over neural networks, linguistics expertise as more important than mathematical expertise, biological expertise as more important than mathematical expertise, and, always, human opinions gathered from a few hundred examples as more valid than statistical tests performed on millions of samples.

When I read about machine learning techniques being applied in the real world, half the time it's by trading firms. I haven't worked for one, so I don't know; but I would bet they are a lot more receptive to new ideas because, unlike scientists, they care about the results more than about their egos. Or at least, an appreciable fraction as much as they care about their egos.

Entry costs

Everybody in science relies on two metrics to decide who to hire and who to give grants to: What their recent publications are, and what school they went to. It is possible to go to a non-top-ranked school and then get on important projects and get publication credits. Someone who just left our company on Friday worked the magic of cranking out good research publications while working as a programmer, always taking on only projects that had good publication potential and never getting stuck with the horrible life-sucking, year-sucking drudgery tasks of, say, converting application X from using database Y to database Z. I just don't know how he did it.

For the most part, that doesn't happen. You don't become a researcher; you start out as a researcher. You need to stay in school, or stay on as a postdoc, until you have your own track record publications and have won your own grant. You need people to read those publications. You don't get to work on important projects and get your work read and get a grant because you're brilliant. You get these things because your advisor works the old-boy network for you. Whatever your field is, there is a network of universities that are recognized as leaders in that field, and you are more-or-less assured of failure in your career (especially in academia or research) unless you go to one of those universities, because you won't get published in good journals, you won't get read much, and you won't get a big grant.

There are exceptions. Fiction writers and computer programmers don't need to go to a fancy university; they need credits and experience. (Computer *programmers.* But don't get a Ph.D. in computer science from a non-elite university and imagine you're going to do research; it won't happen.) Good stories can sort of be recognized; basic knowledge about Enterprise Java can be measured. Companies have recognized the monetary value of doing so. But grant review panels and companies don't really know how to rate scientists or managers, so they try to get somebody from MIT or from Wharton, because nobody ever got fired for buying a Xerox.

The value of scientists to their companies may or may not be reflected in their salaries, but the value of those select universities is certainly reflected in the price of tuition. If your college of choice costs you less than $55,000/yr to attend, including room and board, it will not lead you to success. Unfortunately, the U.S. government won't loan you more than $10,000/yr for tuition.

(One interesting exception is in cosmology. I did a study of successful physicists, as measured by their winning the Nobel or being on the faculty at Harvard. I found that after 1970, no one was successful in physics unless they went to an elite undergraduate college, with a few exceptions. The exceptions were astrophysicists who went to college in Arizona or Hawaii, where there are inexpensive colleges that are recognized as leading institutions in astronomy because they have big telescopes.)

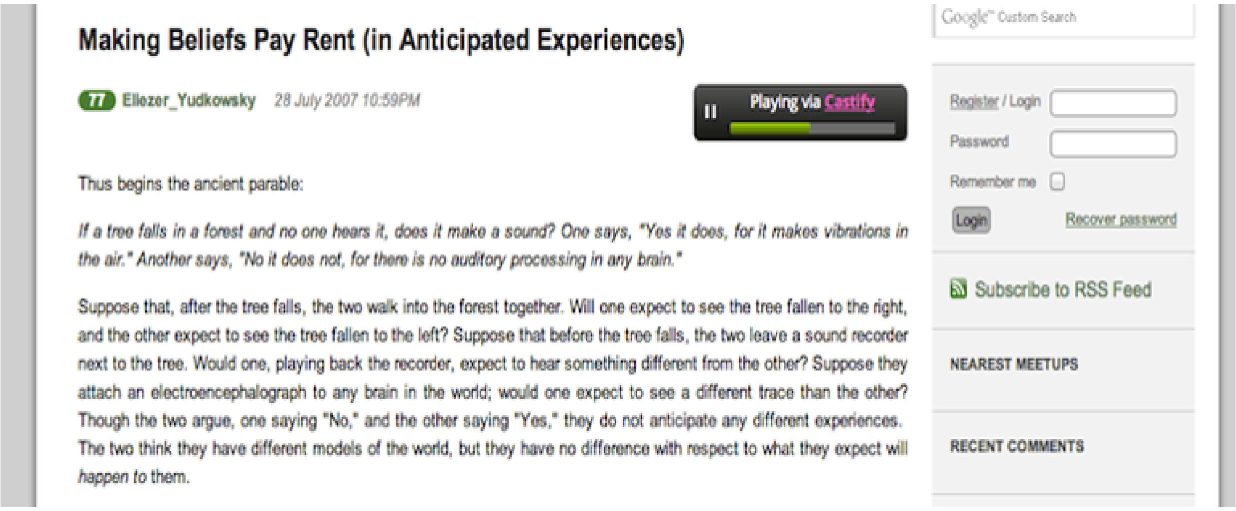
Search

The single-biggest problem with science today is finding relevant results. I have had numerous discussions with experts in a field who were unaware of recent (and not-so-recent) important results in their field because they relied on word-of-mouth and a small set of authoritative journals, while I spent half an hour with Google before our meeting. To take a spectacularly bad example, the literature showing that metronidazole kills Borrelia burgdorferi cysts, while penicillin, doxycycline, amoxicillin, and ceftriaxone do not, is over ten years old; yet metronidazole is never prescribed for Lyme disease while the latter are.

Attention is *the* most-valuable resource in the twenty-first century. Producing a significant result is not hard. Getting people to pay attention to it is. Scientometric analysis of scientific publications shows that producing more and more papers in a field has very little impact on the number of papers cited (a proxy for number of results used), probably because scientists basically read up to one paper per day chosen from one or two leading journals, and that's it. They aren't in the habit of regularly, actively searching for things relevant to their work; and frankly, there isn't much motivation to do that, since using Google to answer a specific question is like using excavation equipment to search for a needle in a haystack.

LessWrong podcasts

Today we're announcing a partnership with Castify to bring you Less Wrong content in audio form. Castify gets blog content read by professional readers and delivers it to their subscribers as a podcast so that you can listen to Less Wrong on the go. The founders of Castify are big fans of Less Wrong so they're rolling out their beta with some of our content.



Making Beliefs Pay Rent (in Anticipated Experiences)

77 Eliezer_Yudkowsky *28 July 2007 10:59PM*

Thus begins the ancient parable:

If a tree falls in a forest and no one hears it, does it make a sound? One says, "Yes it does, for it makes vibrations in the air." Another says, "No it does not, for there is no auditory processing in any brain."

Suppose that, after the tree falls, the two walk into the forest together. Will one expect to see the tree fallen to the right, and the other expect to see the tree fallen to the left? Suppose that before the tree falls, the two leave a sound recorder next to the tree. Would one, playing back the recorder, expect to hear something different from the other? Suppose they attach an electroencephalograph to any brain in the world; would one expect to see a different trace than the other? Though the two argue, one saying "No," and the other saying "Yes," they do not anticipate any different experiences. The two think they have different models of the world, but they have no difference with respect to what they expect will *happen* to them.

Note: The embedded player (above) isn't live as of this posting, but should be deployed soon.

To see how many people will use this, we're having the entire Mysterious Answers to Mysterious Questions core sequence read and recorded. We thought listening to it would be a great way for new readers to get caught up and for others to check out the quality of Castify's work. We will be adding more Less Wrong content based on community feedback, so let us know which content you'd like to see more of in the comments.

For instance: Which other sequences would you like to listen to? Would there be interest in an ongoing podcast channel for the promoted posts?

2012 Survey Results

Thank you to everyone who took the 2012 Less Wrong Survey (the survey is now closed. Do not try to take it.) Below the cut, this post contains the basic survey results, a few more complicated analyses, and the data available for download so you can explore it further on your own. You may want to compare these to the results of the 2011 Less Wrong Survey.

Part 1: Population

How many of us are there?

The short answer is that I don't know.

The 2011 survey ran 33 days and collected 1090 responses. This year's survey ran 23 days and collected 1195 responses. The average number of new responses during the last week was about five per day, so even if I had kept this survey open as long as the last one I probably wouldn't have gotten more than about 1250 responses. That means at most a 15% year on year growth rate, which is pretty abysmal compared to the 650% growth rate in two years we saw last time.

About half of these responses were from lurkers; over half of the non-lurker remainder had commented but never posted to Main or Discussion. That means there were only about 600 non-lurkers.

But I am skeptical of these numbers. I hang out with some people who are very closely associated with the greater Less Wrong community, and a lot of them didn't know about the survey until I mentioned it to them in person. I know some people who could plausibly be described as focusing their lives around the community who just never took the survey for one reason or another. One lesson of this survey may be that the community is no longer limited to people who check Less Wrong very often, if at all. One friend didn't see the survey because she hangs out on the #lesswrong channel more than the main site. Another mostly just goes to meetups. So I think this represents only a small sample of people who could justly be considered Less Wrongers.

The question of "how quickly is LW growing" is also complicated by the high turnover. Over half the people who took this survey said they hadn't participated in the survey last year. I tried to break this down by combining a few sources of information, and I think our 1200 respondents include 500 people who took last year's survey, 400 people who were around last year but didn't take the survey for some reason, and 300 new people.

As expected, there's lower turnover among regulars than among lurkers. Of people who have posted in Main, about 75% took the survey last year; of people who only lurked, about 75% hadn't.

This view of a very high-turnover community and lots of people not taking the survey is consistent with Vladimir Nesov's data showing http://lesswrong.com/lw/e4j/number_of_members_on_lesswrong/77xz 1390 people who have written at least ten comments. But the survey includes only about 600 people who have at least commented; 800ish of Vladimir's accounts are either gone or didn't take the census.

Part 2: Categorical Data

SEX:
Man: 1057, 89.2%
Woman: 120, 10.1%
Other: 2, 0.2%)
No answer: 6, 0.5%

GENDER:
M (cis): 1021, 86.2%
F (cis): 105, 8.9%
M (trans f->m): 3, 0.3%
F (trans m->f): 16, 1.3%
Other: 29, 2.4%
No answer: 11, 0.9%

ORIENTATION:
Heterosexual: 964, 80.7%
Bisexual: 135, 11.4%
Homosexual: 28, 2.4%
Asexual: 24, 2%
Other: 28, 2.4%
No answer: 14, 1.2%

RELATIONSHIP STYLE:
Prefer monogamous: 639, 53.9%
Prefer polyamorous: 155, 13.1%
Uncertain/no preference: 358, 30.2%
Other: 21, 1.8%
No answer: 12, 1%

NUMBER OF CURRENT PARTNERS:
0: 591, 49.8%
1: 519, 43.8%
2: 34, 2.9%
3: 12, 1%
4: 5, 0.4%
6: 1, 0.1%
7, 1, 0.1% *(and this person added "really, not trolling")*
Confusing or no answer: 20, 1.8%

RELATIONSHIP STATUS:
Single: 628, 53%
Relationship: 323, 27.3%
Married: 220, 18.6%
No answer: 14, 1.2%

RELATIONSHIP GOALS:
Not looking for more partners: 707, 59.7%
Looking for more partners: 458, 38.6%
No answer: 20, 1.7%

COUNTRY:
USA: 651, 54.9%
UK: 103, 8.7%
Canada: 74, 6.2%
Australia: 59, 5%
Germany: 54, 4.6%
Israel: 15, 1.3%
Finland: 15, 1.3%
Russia: 13, 1.1%
Poland: 12, 1%

These are all the countries with greater than 1% of Less Wrongers, but other, more exotic locales included Kenya, Pakistan, and Iceland, with one user each. You can see the full table here.

This data also allows us to calculate Less Wrongers per capita:

Finland: 1/366,666
Australia: 1/389,830
Canada: 1/472,972
USA: 1/483,870
Israel: 1/533,333
UK: 1/603,883
Germany: 1/1,518,518
Poland: 1/3,166,666
Russia: 1/11,538,462

RACE:
White, non-Hispanic 1003, 84.6%
East Asian: 50, 4.2%
Hispanic 47, 4.0%
Indian Subcontinental 28, 2.4%
Black 8, 0.7%
Middle Eastern 4, 0.3%
Other: 33, 2.8%
No answer: 12, 1%

WORK STATUS:
Student: 476, 40.7%
For-profit work: 364, 30.7%
Self-employed: 95, 8%
Unemployed: 81, 6.8%
Academics (teaching): 54, 4.6%
Government: 46, 3.9%
Non-profit: 44, 3.7%
Independently wealthy: 12, 1%
No answer: 13, 1.1%

PROFESSION:
Computers (practical): 344, 29%
Math: 109, 9.2%
Engineering: 98, 8.3%
Computers (academic): 72, 6.1%

Physics: 66, 5.6%
Finance/Econ: 65, 5.5%
Computers (AI): 39, 3.3%
Philosophy: 36, 3%
Psychology: 25, 2.1%
Business: 23, 1.9%
Art: 22, 1.9%
Law: 21, 1.8%
Neuroscience: 19, 1.6%
Medicine: 15, 1.3%
Other social science: 24, 2%
Other hard science: 20, 1.7%
Other: 123, 10.4%
No answer: 27, 2.3%

DEGREE:
Bachelor's: 438, 37%
High school: 333, 28.1%
Master's: 192, 16.2%
Ph.D: 71, 6%
2-year: 43, 3.6%
MD/JD/professional: 24, 2%
None: 55, 4.6%
Other: 15, 1.3%
No answer: 14, 1.2%

POLITICS:
Liberal: 427, 36%
Libertarian: 359, 30.3%
Socialist: 326, 27.5%
Conservative: 35, 3%
Communist: 8, 0.7%
No answer: 30, 2.5%

You can see the exact definitions given for each of these terms on the survey.

RELIGIOUS VIEWS:
Atheist, not spiritual: 880, 74.3%
Atheist, spiritual: 107, 9.0%
Agnostic: 94, 7.9%
Committed theist: 37, 3.1%
Lukewarm theist: 27, 2.3%
Deist/Pantheist/etc: 23, 1.9%
No answer: 17, 1.4%

FAMILY RELIGIOUS VIEWS:
Lukewarm theist: 392, 33.1%
Committed theist: 307, 25.9%
Atheist, not spiritual: 161, 13.6
Agnostic: 149, 12.6%
Atheist, spiritual: 46, 3.9%
Deist/Pantheist/Etc: 32, 2.7%
Other: 84, 7.1%

RELIGIOUS BACKGROUND:
Other Christian: 517, 43.6%
Catholic: 295, 24.9%
Jewish: 100, 8.4%
Hindu: 21, 1.8%
Traditional Chinese: 17, 1.4%
Mormon: 15, 1.3%
Muslim: 12, 1%

Raw data is available here.

MORAL VIEWS:
Consequentialism: 735, 62%
Virtue Ethics: 166, 14%
Deontology: 50, 4.2%
Other: 214, 18.1%
No answer: 20, 1.7%

NUMBER OF CHILDREN
0: 1044, 88.1%
1: 51, 4.3%
2: 48, 4.1%
3: 19, 1.6%
4: 3, 0.3%
5: 2, 0.2%
6: 1, 0.1%
No answer: 17, 1.4%

WANT MORE CHILDREN?
No: 438, 37%
Maybe: 363, 30.7%
Yes: 366, 30.9%
No answer: 16, 1.4%

LESS WRONG USE:
Lurkers (no account): 407, 34.4%
Lurkers (with account): 138, 11.7%
Posters (comments only): 356, 30.1%
Posters (comments + Discussion only): 164, 13.9%
Posters (including Main): 102, 8.6%

SEQUENCES:
Never knew they existed until this moment: 99, 8.4%
Knew they existed; never looked at them: 23, 1.9%
Read < 25%: 227, 19.2%
Read ~ 25%: 145, 12.3%
Read ~ 50%: 164, 13.9%
Read ~ 75%: 203, 17.2%
Read ~ all: 306, 24.9%
No answer: 16, 1.4%

Dear 8.4% of people: there is this collection of old blog posts called the Sequences. It is by Eliezer, the same guy who wrote Harry Potter and the Methods of Rationality. It is really good! If you read it, you will understand what we're talking about much better!

REFERRALS:
Been here since Overcoming Bias: 265, 22.4%
Referred by a link on another blog: 23.5%
Referred by a friend: 147, 12.4%
Referred by HPMOR: 262, 22.1%
No answer: 35, 3%

BLOG REFERRALS:
Common Sense Atheism: 20 people
Hacker News: 20 people
Reddit: 15 people
Unequally Yoked: 7 people
TV Tropes: 7 people
Marginal Revolution: 6 people
gwern.net: 5 people
RationalWiki: 4 people
Shtetl-Optimized: 4 people
XKCD fora: 3 people
Accelerating Future: 3 people

These are all the sites that referred at least three people in a way that was obvious to disentangle from the raw data. You can see a more complete list, including the long tail, here.

MEETUPS:
Never been to one: 834, 70.5%
Have been to one: 320, 27%
No answer: 29, 2.5%

CATASTROPHE:
Pandemic (bioengineered): 272, 23%
Environmental collapse: 171, 14.5%
Unfriendly AI: 160, 13.5%
Nuclear war: 155, 13.1%
Economic/Political collapse: 137, 11.6%
Pandemic (natural): 99, 8.4%
Nanotech: 49, 4.1%
Asteroid: 43, 3.6%

The wording of this question was "which disaster do you think is most likely to wipe out greater than 90% of humanity before the year 2100?"

CRYONICS STATUS:
No, don't want to: 275, 23.2%
No, still thinking: 472, 39.9%
No, procrastinating: 178, 15%
No, unavailable: 120, 10.1%
Yes, signed up: 44, 3.7%
Never thought about it: 46, 3.9%
No answer: 48, 4.1%

VEGETARIAN:
No: 906, 76.6%

Yes: 147, 12.4%
No answer: 130, 11%

For comparison, 3.2% of US adults are vegetarian.

SPACED REPETITION SYSTEMS
Don't use them: 511, 43.2%
Do use them: 235, 19.9%
Never heard of them: 302, 25.5%

Dear 25.5% of people: spaced repetition systems are nifty, mostly free computer programs that allow you to study and memorize facts more efficiently. See for example http://ankisrs.net/

HPMOR:
Never read it: 219, 18.5%
Started, haven't finished: 190, 16.1%
Read all of it so far: 659, 55.7%

Dear 18.5% of people: Harry Potter and the Methods of Rationality is a Harry Potter fanfic about rational thinking written by Eliezer Yudkowsky (the guy who started this site). It's really good. You can find it at http://www.hpmor.com/.

ALTERNATIVE POLITICS QUESTION:
Progressive: 429, 36.3%
Libertarian: 278, 23.5%
Reactionary: 30, 2.5%
Conservative: 24, 2%
Communist: 22, 1.9%
Other: 156, 13.2%

ALTERNATIVE ALTERNATIVE POLITICS QUESTION:
Left-Libertarian: 102, 8.6%
Progressive: 98, 8.3%
Libertarian: 91, 7.7%
Pragmatist: 85, 7.2%
Social Democrat: 80, 6.8%
Socialist: 66, 5.6%
Anarchist: 50, 4.1%
Futarchist: 29, 2.5%
Moderate: 18, 1.5%
Moldbuggian: 19, 1.6%
Objectivist: 11, 0.9%

These are the only ones that had more than ten people. Other responses notable for their unusualness were Monarchist (5 people), fascist (3 people, plus one who was up for fascism but only if he could be the leader), conservative (9 people), and a bunch of people telling me politics was stupid and I should feel bad for asking the question. You can see the full table here.

CAFFEINE:
Never: 162, 13.7%
Rarely: 237, 20%
At least 1x/week: 207, 17.5

Daily: 448, 37.9
No answer: 129, 10.9%

SMOKING:
Never: 896, 75.7%
Used to: 1-5, 8.9%
Still do: 51, 4.3%
No answer: 131, 11.1%

For comparison, about 28.4% of the US adult population smokes

NICOTINE (OTHER THAN SMOKING):
Never used: 916, 77.4%
Rarely use: 82, 6.9%
>1x/month: 32, 2.7%
Every day: 14, 1.2%
No answer: 139, 11.7%

MODAFINIL:
Never: 76.5%
Rarely: 78, 6.6%
>1x/month: 48, 4.1%
Every day: 9, 0.8%
No answer: 143, 12.1%

TRUE PRISONERS' DILEMMA:
Defect: 341, 28.8%
Cooperate: 316, 26.7%
Not sure: 297, 25.1%
No answer: 229, 19.4%

FREE WILL:
Not confused: 655, 55.4%
Somewhat confused: 296, 25%
Confused: 81, 6.8%
No answer: 151, 12.8%

TORTURE VS. DUST SPECKS
Choose dust specks: 435, 36.8%
Choose torture: 261, 22.1%
Not sure: 225, 19%
Don't understand: 22, 1.9%
No answer: 240, 20.3%

SCHRODINGER EQUATION:
Can't calculate it: 855, 72.3%
Can calculate it: 175, 14.8%
No answer: 153, 12.9%

PRIMARY LANGUAGE:
English: 797, 67.3%
German: 54, 4.5%
French: 13, 1.1%
Finnish: 11, 0.9%

Dutch: 10, 0.9%
Russian: 15, 1.3%
Portuguese: 10, 0.9%

These are all the languages with ten or more speakers, but we also have everything from Marathi to Tibetan. You can see the full table here..

NEWCOMB'S PROBLEM

One-box: 726, 61.4%
Two-box: 78, 6.6%
Not sure: 53, 4.5%
Don't understand: 86, 7.3%
No answer: 240, 20.3%

ENTREPRENEUR:

Don't want to start business: 447, 37.8%
Considering starting business: 334, 28.2%
Planning to start business: 96, 8.1%
Already started business: 112, 9.5%
No answer: 194, 16.4%

ANONYMITY:

Post using real name: 213, 18%
Easy to find real name: 256, 21.6%
Hard to find name, but wouldn't bother me if someone did: 310, 26.2%
Anonymity is very important: 170, 14.4%
No answer: 234, 19.8%

HAVE YOU TAKEN A PREVIOUS LW SURVEY?

No: 559, 47.3%
Yes: 458, 38.7%
No answer: 116, 14%

TROLL TOLL POLICY:

Disapprove: 194, 16.4%
Approve: 178, 15%
Haven't heard of this: 375, 31.7%
No opinion: 249, 21%
No answer: 187, 15.8%

MYERS-BRIGGS

INTJ: 163, 13.8%
INTP: 143, 12.1%
ENTJ: 35, 3%
ENTP: 30, 2.5%
INFP: 26, 2.2%
INFJ: 25. 2.1%
ISTJ: 14, 1.2%
No answer: 715, 60%

This includes all types with greater than 10 people. You can see the full table here.

Part 3: Numerical Data

Except where indicated otherwise, all the numbers below are given in the format:

mean+standard_deviation (25% level, 50% level/median, 75% level) [n = number of data points]

INTELLIGENCE:

IQ (self-reported): 138.7 + 12.7 (130, 138, 145) [n = 382]
SAT (out of 1600): 1485.8 + 105.9 (1439, 1510, 1570) [n = 321]
SAT (out of 2400): 2319.5 + 1433.7 (2155, 2240, 2320)
ACT: 32.7 + 2.3 (31, 33, 34) [n = 207]
IQ (on iqtest.dk): 125.63 + 13.4 (118, 130, 133)   [n = 378]

I am going to harp on these numbers because in the past some people have been pretty quick to ridicule this survey's intelligence numbers as completely useless and impossible and so on.

According to IQ Comparison Site, an SAT score of 1485/1600 corresponds to an IQ of about 144. According to Ivy West, an ACT of 33 corresponds to an SAT of 1470 (and thence to IQ of 143).

So if we consider self-report, SAT, ACT, and iqtest.dk as four measures of IQ, these come out to 139, 144, 143, and 126, respectively.

All of these are pretty close except iqtest.dk. I ran a correlation between all of them and found that self-reported IQ is correlated with SAT scores at the 1% level and iqtest.dk at the 5% level, but SAT scores and IQTest.dk are not correlated with each other.

Of all these, I am least likely to trust iqtest.dk. First, it's a random Internet IQ test. Second, it correlates poorly with the other measures. Third, a lot of people have complained in the comments to the survey post that it exhibits some weird behavior.

But iqtest.dk gave us the lowest number! And even it said the average was 125 to 130! So I suggest that we now have pretty good, pretty believable evidence that the average IQ for this site really is somewhere in the 130s, and that self-reported IQ isn't as terrible a measure as one might think.

AGE:
27.8 + 9.2 (22, 26, 31) [n = 1185]

LESS WRONG USE:
Karma: 1078 + 2939.5 (0, 4.5, 136) [n = 1078]
Months on LW: 26.7 + 20.1 (12, 24, 40) [n = 1070]
Minutes/day on LW: 19.05 + 24.1 (5, 10, 20) [n = 1105]
Wiki views/month: 3.6 + 6.3 (0, 1, 5) [n = 984]
Wiki edits/month: 0.1 + 0.8 (0, 0, 0) [n = 984]

PROBABILITIES:
Many Worlds: 51.6 + 31.2 (25, 55, 80) [n = 1005]

Aliens (universe): 74.2 + 32.6 (50, 90, 99) [n = 1090]
Aliens (galaxy): 42.1 + 38 (5, 33, 80) [n = 1081]
Supernatural: 5.9 + 18.6 (0, 0, 1) [n = 1095]
God: 6 + 18.7 (0, 0, 1) [n = 1098]
Religion: 3.8 + 15.5 (0, 0, 0.8) [n = 1113]
Cryonics: 18.5 + 24.8 (2, 8, 25) [n = 1100]
Antiagathics: 25.1 + 28.6 (1, 10, 35) [n = 1094]
Simulation: 25.1 + 29.7 (1, 10, 50) [n = 1039]
Global warming: 79.1 + 25 (75, 90, 97) [n = 1112]
No catastrophic risk: 71.1 + 25.5 (55, 80, 90) [n = 1095]
Space: 20.1 + 27.5 (1, 5, 30) [n = 953]

CALIBRATION:
Year of Bayes' birth: 1767.5 + 109.1 (1710, 1780, 1830) [n = 1105]
Confidence: 33.6 + 23.6 (20, 30, 50) [n= 1082]

MONEY:
Income/year: 50,913 + 60644.6 (12000, 35000, 74750) [n = 644]
Charity/year: 444.1 + 1152.4 (0, 30, 250) [n = 950]
SIAI/CFAR charity/year: 309.3 + 3921 (0, 0, 0) [n = 961]
Aging charity/year: 13 + 184.9 (0, 0, 0) [n = 953]

TIME USE:
Hours online/week: 42.4 + 30 (21, 40, 59) [n = 944]
Hours reading/week: 30.8 + 19.6 (18, 28, 40) [n = 957]
Hours writing/week: 7.9 + 9.8 (2, 5, 10) [n = 951]

POLITICAL COMPASS:
Left/Right: -2.4 + 4 (-5.5, -3.4, -0.3) [n = 476]
Libertarian/Authoritarian: -5 + 2 (-6.2, -5.2, -4)

BIG 5 PERSONALITY TEST:
Big 5 (O): 60.6 + 25.7 (41, 65, 84) [n = 453]
Big 5 (C): 35.2 + 27.5 (10, 30, 58) [n = 453]
Big 5 (E): 30.3 + 26.7 (7, 22, 48) [n = 454]
Big 5 (A): 41 + 28.3 (17, 38, 63) [n = 453]
Big 5 (N): 36.6 + 29 (11, 27, 60) [n = 449]

These scores are in percentiles, so LWers are more Open, but less Conscientious, Agreeable, Extraverted, and Neurotic than average test-takers. Note that people who take online psychometric tests are probably a pretty skewed category already so this tells us nothing. Also, several people got confusing results on this test or found it different than other tests that they took, and I am pretty unsatisfied with it and don't trust the results.

AUTISM QUOTIENT
AQ: 24.1 + 12.2 (17, 24, 30) [n = 367]

This test says the average control subject got 16.4 and 80% of those diagnosed with autism spectrum disorders get 32+ (which of course doesn't tell us what percent of people above 32 have autism...). If we trust them, most LWers are more autistic than average.

CALIBRATION:

Reverend Thomas Bayes was born in 1701. Survey takers were asked to guess this date within 20 years, so anyone who guessed between 1681 and 1721 was recorded as getting a correct answer. The percent of people who answered correctly is recorded below, stratified by the confidence they gave of having guessed correctly and with the number of people at that confidence level.

0-5: 10% [n = 30]
5-15: 14.8% [n = 183]
15-25: 10.3% [n = 242]
25-35: 10.7% [n = 225]
35-45: 11.2% [n = 98]
45-55: 17% [n = 118]
55-65: 20.1% [n = 62]
65-75: 26.4% [n = 34]
75-85: 36.4% [n = 33]
85-95: 60.2% [n = 20]
95-100: 85.7% [n = 23]

Here's a classic calibration chart. The blue line is perfect calibration. The orange line is you guys. And the yellow line is average calibration from an experiment I did with untrained subjects a few years ago (which of course was based on different questions and so not directly comparable).

The results are atrocious; when Less Wrongers are 50% certain, they only have about a 17% chance of being correct. On this problem, at least, they are as bad or worse at avoiding overconfidence bias as the general population.

My hope was that this was the result of a lot of lurkers who don't know what they're doing stumbling upon the survey and making everyone else look bad, so I ran a second analysis. This one used only the numbers of people who had been in the community at least 2 years and accumulated at least 100 karma; this limited my sample size to about 210 people.

I'm not going to post exact results, because I made some minor mistakes which means they're off by a percentage point or two, but the general trend was that they looked exactly like the results above: atrocious. If there is some core of elites who are less biased than the general population, they are well past the 100 karma point and probably too rare to feel confident even detecting at this kind of a sample size.

I really have no idea what went so wrong. Last year's results were pretty good - encouraging, even. I wonder if it's just an especially bad question. Bayesian statistics is pretty new; one would expect Bayes to have been born in rather more modern times. It's also possible that I've handled the statistics wrong on this one; I wouldn't mind someone double-checking my work.

Or we could just be *really horrible*. If we haven't even learned to avoid the one bias that we can measure super well and which is most susceptible to training, what are we even *doing* here? Some remedial time at PredictionBook might be in order.

HYPOTHESIS TESTING:

I tested a very few of the possible hypothesis that were proposed in the survey design threads.

Are people who understand quantum mechanics are more likely to believe in Many Worlds? We perform a t-test, checking whether one's probability of the MWI being true depends on whether or not one can solve the Schrodinger Equation. People who could solve the equation had on average a 54.3% probability of MWI, compared to 51.3% in those who could not. The p-value is 0.26; there is a 26% probability this occurs by chance. Therefore, we fail to establish that people's probability of MWI varies with understanding of quantum mechanics.

Are there any interesting biological correlates of IQ? We run a correlation between self-reported IQ, height, maternal age, and paternal age. The correlations are in the expected direction but not significant.

Are there differences in the ways men and women interact with the community? I had sort of vaguely gotten the impression that women were proportionally younger, newer to the community, and more likely to be referred via HPMOR. The average age of women on LW is 27.6 compared to 27.7 for men; obviously this difference is not significant. 14% of the people referred via HPMOR were women compared to about 10% of the community at large, but this difference is pretty minor. Women were on average newer to the community - 21 months vs. 39 for men - but to my surprise a t-test was unable to declare this significant. Maybe I'm doing it wrong?

Does the amount of time spent in the community affect one's beliefs in the same way as in previous surveys? I ran some correlations and found that it does. People who have been around longer continue to be more likely to believe in MWI, less likely to believe in aliens in the universe (though not in our galaxy), and less likely to believe in God (though not religion). There was no effect on cryonics this time.

In addition, the classic correlations between different beliefs continue to hold true. There is an obvious cluster of God, religion, and the supernatural. There's also a scifi cluster of cryonics, antiagathics, MWI, aliens, and the Simulation Hypothesis, and catastrophic risk (this also seems to include global warming, for some reason).

Are there any differences between men and women in regards to their belief in these clusters? We run a t-test between men and women. Men and women have about the same probability of God (men: 5.9, women: 6.2, p = .86) and similar results for the rest of the religion cluster, but men have much higher beliefs in for example antiagathics (men 24.3, women: 10.5, p < .001) and the rest of the scifi cluster.

DESCRIPTIONS OF LESS WRONG

Survey users were asked to submit a description of Less Wrong in 140 characters or less. I'm not going to post all of them, but here is a representative sample:

- "Probably the most sensible philosophical resource avaialble."
- "Contains the great Sequences, some of Luke's posts, and very little else."
- "The currently most interesting site I found ont the net."
- "EY cult"
- "How to think correctly, precisely, and efficiently."
- "HN for even bigger nerds."
- "Social skills philosophy and AI theorists on the same site, not noticing each other."
- "Cool place. Any others like it?"

- "How to avoid predictable pitfalls in human psychology, and understand hard things well: The Website."
- "A bunch of people trying to make sense of the wold through their own lens, which happens to be one of calculation and rigor"
- "Nice."
- "A font of brilliant and unconventional wisdom."
- "One of the few sane places on Earth."
- "Robot god apocalypse cult spinoff from Harry Potter."
- "A place to converse with intelligent, reasonably open-minded people."
- "Callahan's Crosstime Saloon"
- "Amazing rational transhumanist calming addicting Super Reddit"
- "Still wrong"
- "A forum for helping to train people to be more rational"
- "A very bright community interested in amateur ethical philosophy, mathematics, and decision theory."
- "Dying. Social games and bullshit now >50% of LW content."
- "The good kind of strange, addictive, so much to read!"
- "Part genuinely useful, part mental masturbation."
- "Mostly very bright and starry-eyed adults who never quite grew out of their science-fiction addiction as adolescents."
- "Less Wrong: Saving the world with MIND POWERS!"
- "Perfectly patternmatches the 'young-people-with-all-the-answers' cliche"
- "Rationalist community dedicated to self-improvement."
- "Sperglord hipsters pretending that being a sperglord hipster is cool." *(this person's Autism Quotient was two points higher than LW average, by the way)*
- "An interesting perspective and valuable database of mental techniques."
- "A website with kernels of information hidden among aspy nonsense."
- "Exclusive, elitist, interesting, potentially useful, personal depression trigger."
- "A group blog about rationality and related topics. Tends to be overzealous about cryogenics and other pet ideas of Eliezer Yudkowsky."
- "Things to read to make you think better."
- "Excellent rationality. New-age self-help. Worrying groupthink."
- "Not a cult at all."
- "A cult."
- "The new thing for people who would have been Randian Objectivists 30 years ago."
- "Fascinating, well-started, risking bloat and failure modes, best as archive."
- "A fun, insightful discussion of probability theory and cognition."
- "More interesting than useful."
- "The most productive and accessible mind-fuckery on the Internet."
- "A blog for rationality, cognitive bias, futurism, and the Singularity."
- "Robo-Protestants attempting natural theology."
- "Orderly quagmire of tantalizing ideas drawn from disagreeable priors."
- "Analyze everything. And I do mean everything. Including analysis. Especially analysis. And analysis of analysis."
- "Very interesting and sometimes useful."
- "Where people discuss and try to implement ways that humans can make their values, actions, and beliefs more internally consistent."
- "Eliezer Yudkowsky personality cult."
- "It's like the Mormons would be if everyone were an atheist and good at math and didn't abstain from substances."
- "Seems wacky at first, but gradually begins to seem normal."
- "A varied group of people interested in philosophy with high Openness and a methodical yet amateur approach."
- "Less Wrong is where human algorithms go to debug themselves."

- "They're kind of like a cult, but that doesn't make them wrong."
- "A community blog devoted to nerds who think they're smarter than everyone else."
- "90% sane! A new record!"
- "The Sequences are great. LW now slowly degenerating to just another science forum."
- "The meetup groups are where it's at, it seems to me. I reserve judgment till I attend one."
- "All I really know about it is this long survey I took."
- "The royal road of rationality."
- "Technically correct: The best kind of correct!"
- "Full of angry privilege."
- "A sinister instrument of billionaire Peter Thiel."
- "Dangerous apocalypse cult bent on the systematic erasure of traditional values and culture by any means necessary."
- "Often interesting, but I never feel at home."
- "One of the few places I truly feel at home, knowing that there are more people like me."
- "Currently the best internet source of information-dense material regarding cog sci, debiasing, and existential risk."
- "Prolific and erudite writing on practical techniques to enhance the effectiveness of our reason."
- "An embarrassing Internet community formed around some genuinely great blog writings."
- "I bookmarked it a while ago and completely forgot what it is about. I am taking the survey to while away my insomnia."
- "A somewhat intimidating but really interesting website that helps refine rational thinking."
- "A great collection of ways to avoid systematic bias and come to true and useful conclusions."
- "Obnoxious self-serving, foolish trolling dehumanizing pseudointellectualism, aesthetically bankrupt."
- "The cutting edge of human rationality."
- "A purveyor of exceedingly long surveys."

PUBLIC RELEASE

That last commenter was right. This survey had vastly more data than any previous incarnation; although there are many more analyses I would like to run I am pretty exhausted and I know people are anxious for the results. I'm going to let CFAR analyze and report on their questions, but the rest should be a community effort. So I'm releasing the survey to everyone in the hopes of getting more information out of it. If you find something interesting you can either post it in the comments or start a new thread somewhere.

The data I'm providing is the raw data EXCEPT:

- I deleted a few categories that I removed halfway through the survey for various reasons
- I deleted 9 entries that were duplicates of other entries, ie someone pressed 'submit' twice.
- I deleted the timestamp, which would have made people extra-identifiable, and sorted people by their CFAR random number to remove time order information.
- I removed one person whose information all came out as weird symbols.
- I numeralized some of the non-numeric data, especially on the number of months in

community question. This is not the version I cleaned up fully, so you will get to experience some of the same pleasure I did working with the rest.
- I deleted 117 people who either didn't answer the privacy question or who asked me to keep them anonymous, leaving 1067 people.

Here it is: **Data in .csv format** , **Data in Excel format**

The Relation Projection Fallacy and the purpose of life

I bet most people here have realized this explicitly or implicitly, but this comment has inspired me to write a short, linkable summary of this error pattern, with a name:

The Relation Projection Fallacy: a denotational error whereby one confuses an n-ary relation for an m-ary relation, where usually m<n.

Example instance: "Life has no purpose."

This is a troublesome phrase. Why? If you look at unobjectionable uses of the concept <purpose> --- also referenced by synonyms like "having a point" --- it is in fact a **ternary relation.**

Example non-instance: "The purpose of a doorstop is to stop doors."

Here, one can query "to whom?" and be returned the context "to the person who made it" or "to the person who's using it", etc. That is, the full denotation of "purpose" is always of the form "The purpose of X to Y is Z," where Y is often implicit or can take a wide range of values.

This has nothing to do with connotation... it's just how the concept <purpose> typically works as people use it. But to flog a dead horse, the purpose of a doorstop to a cat may be to make an amusing sound as it glides across the floor after the cat hits it. The value of Y always matters. There is no "true purpose" stored anywhere inside the doorstop, or even in the combination of the doorstop and the door it is stopping. To think otherwise is literally *projecting, in the mathematical sense,* a ternary relation, i.e., a subset of a product of three sets (objects)x(agents)x(verbs), into a product of two sets, (objects)x(verbs). But people often do this projection incorrectly, by either searching for a purpose that is intrinsic to the Doorstop or to Life, or by searching for a canonical value of "Y" like "The Great Arbiter of Purpose", both of which are not to be found, at least to their satisfaction when they utter the phrase "Life has no purpose."

Likewise, the relation "has a purpose" is typically a **binary relation**, because again, we can always ask "to whom?". "<That doorstop> has a purpose to <me>."

In some form, this realization is of course the cause of many schools of thought taking the name "relativist" on many different issues. But I find that people over-use the phrase "It's all relative" to connote "It's all meaningless" or "there is no answer". Which is ironic, because meaning itself is a ternary relation! Its typical denotation is of the form "The meaning of X to Y is Z", like in

- "The meaning of <the sound 'owe'> to <French people> is <liquid water>" or
- "The meaning of <that pendant> to <your mother> is <a certain undescribed experience of sentimentality>".

Realizing this should NOT result in a cascade of bottomless relativism where nothing means anything! In fact, the first time I had this thought as a kid, I arrived at the connotationally pleasing conclusion "My life can have as many purposes as there are agents for it to have a purpose to."

Indeed, the meaning of <"purpose"> to <humans> is <a certain ternary functional relationship between objects, agents, and verbs>, and the meaning of <"meaning"> to <humans> is <a certain ternary relationship between syntactic elements, people generating or perceiving them, and referents>.

When I found LessWrong, I was happy to find that Eliezer wrote on almost exactly this realization in 2-Place and 1-Place Words, but sad that the post had few upvotes -- only 14 right now. So in case it was too long, or didn't have a snappy enough name, I thought I'd try giving the idea another shot.

ETA: In the special case of talking to someone wondering about the purpose of *life*, here is how I would use this observation *in the form of an argument:*

> First of all, you may be lacking satisfaction in your life for some reason, and framing this to yourself in philosophical terms like "Life has no purpose, because <argument>." If that's true, it's quite likely that you'd feel differently if your emotional needs as a social primate were being met, and in that sense the solution is not an "answer" but rather some actions that will result in these needs being met.
>
> Still, that does not address the <argument>. So because "What is s the purpose of life?" may be a hard question, let's look at easier examples of *purpose* and see how they work. Notice how they all have someone the purpose is *to*? And how that's missing in your "purpose of life" question? Because of that, you could end up feeling one of two ways:
>
> (1) Satisfied, because now you can just ask "What could be the purpose of my life to <my friends, my family, myself, the world at large, etc>", and come up with answers, or
>
> (2) Unsatisfied, because there is no agent to ask about such that the answer would seem important enough to you.
>
> And I claim that whether you end up at (1) or (2) is probably more a function of whether your social primate emotional needs are being met than any particular philosophical argument.

That being said, if you believe this argument, the best thing to do for someone lacking a sense of purpose is probably not to just say the argument, but to help them start satisfying their emotional needs, and have this argument mainly to satisfy their sense of curiosity or nagging intellectual doubts about the issue.

Train Philosophers with Pearl and Kahneman, not Plato and Kant

Part of the sequence: Rationality and Philosophy

> Hitherto the people attracted to philosophy have been mostly those who loved the big generalizations, which were all wrong, so that few people with exact minds have taken up the subject.
>
> Bertrand Russell

I've complained before that philosophy is a diseased discipline which spends far too much of its time debating definitions, ignoring relevant scientific results, and endlessly re-interpreting old dead guys who didn't know the slightest bit of 20th century science. Is that *still* the case?

You bet. There's *some* good philosophy out there, but much of it is bad enough to make CMU philosopher Clark Glymour suggest that on tight university budgets, philosophy departments could be defunded unless their work is useful to (cited by) scientists and engineers — just as his own work on causal Bayes nets is now widely used in artificial intelligence and other fields.

How did philosophy get this way? Russell's hypothesis is not too shabby. Check the syllabi of the undergraduate "intro to philosophy" classes at the world's top 5 U.S. philosophy departments — NYU, Rutgers, Princeton, Michigan Ann Arbor, and Harvard — and you'll find that they spend a *lot* of time with (1) old dead guys who were wrong about almost everything because they knew nothing of modern logic, probability theory, or science, and with (2) 20th century philosophers who were way too enamored with cogsci-ignorant armchair philosophy. (I say more about the reasons for philosophy's degenerate state here.)

As the CEO of a philosophy/math/compsci research institute, I think many philosophical problems are important. But the field of philosophy doesn't seem to be very good at answering them. What can we do?

Why, come up with better philosophical methods, of course!

Scientific methods have improved over time, and so can philosophical methods. Here is the *first* of my recommendations...

More Pearl and Kahneman, less Plato and Kant

Philosophical training should begin with the latest and greatest formal methods ("Pearl" for the probabilistic graphical models made famous in Pearl 1988), and the latest and greatest science ("Kahneman" for the science of human reasoning reviewed in Kahneman 2011). Beginning with Plato and Kant (and company), as most universities do today, both (1) filters for inexact thinkers, as Russell suggested, and

(2) teaches people to have too much respect for failed philosophical methods that are out of touch with 20th century breakthroughs in math and science.

So, I recommend we teach young philosophy students:

more Bayesian rationality, heuristics and biases, & debiasing, *less* informal "critical thinking skills";
more mathematical logic & theory of computation, *less* term logic;
more probability theory & Bayesian scientific method, *less* pre-1980 philosophy of science;
more psychology of concepts & machine learning, *less* conceptual analysis;
more formal epistemology & computational epistemology, *less* pre-1980 epistemology;
more physics & cosmology, *less* pre-1980 metaphysics;
more psychology of choice, *less* philosophy of free will;
more moral psychology, decision theory, and game theory, *less* intuitionist moral philosophy;
more cognitive psychology & cognitive neuroscience, *less* pre-1980 philosophy of mind;
more linguistics & psycholinguistics, *less* pre-1980 philosophy of language;
more neuroaesthetics, *less* aesthetics;
more causal models & psychology of causal perception, *less* pre-1980 theories of causation.

(In other words: train philosophy students like they do at CMU, but even "more so.")

So, my own "intro to philosophy" mega-course might be guided by the following core readings:

1. Stanovich, *Rationality and the Reflective Mind* (2010)
2. Hinman, *Fundamentals of Mathematical Logic* (2005)
3. Russell & Norvig, *Artificial Intelligence: A Modern Approach* (3rd edition, 2009) — contains chapters which briefly introduce probability theory, probabilistic graphical models, computational decision theory and game theory, knowledge representation, machine learning, computational epistemology, and other useful subjects
4. Sipser, *Introduction to the Theory of Computation* (3rd edition, 2012) — relevant to lots of philosophical problems, as discussed in Aaronson (2011)
5. Howson & Urbach, *Scientific Reasoning: The Bayesian Approach* (3rd edition, 2005)
6. Holyoak & Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning* (2012) — contains chapters which briefly introduce the psychology of knowledge representation, concepts, categories, causal learning, explanation, argument, decision making, judgment heuristics, moral judgment, behavioral game theory, problem solving, creativity, and other useful subjects
7. Dolan & Sharot (eds.), *Neuroscience of Preference and Choice* (2011)
8. Krane, *Modern Physics* (3rd edition, 2012) — includes a brief introduction to cosmology

(There are many prerequisites to these, of course. I think philosophy should be a Highly Advanced subject of study that requires lots of prior training in maths and the sciences, like string theory but hopefully more productive.)

Once students are equipped with some of the latest math and science, *then* let them tackle The Big Questions. I bet they'd get farther than those raised on Plato and Kant instead.

You might also let them read 20th century analytic philosophy at that point — hopefully their training will have inoculated them from picking up bad thinking habits.

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

Against NHST

> A summary of standard non-Bayesian criticisms of common frequentist statistical practices, with pointers into the academic literature.

Frequentist statistics is a wide field, but in practice by innumerable psychologists, biologists, economists etc, frequentism tends to be a particular style called "Null Hypothesis Significance Testing" (NHST) descended from R.A. Fisher (as opposed to eg. Neyman-Pearson) which is focused on

1. setting up a null hypothesis and an alternative hypothesis
2. calculating a *p*-value (possibly via a _<_a href="https://en.wikipedia.org/wiki/Student%27s_t-test">t-test or more complex alternatives like ANOVA)
3. and rejecting the null if an arbitrary threshold is passed.

NHST became nearly universal between the 1940s & 1960s (see Gigerenzer 2004, pg18), and has been heavily criticized for as long. Frequentists criticize it for:

1. practitioners & statistics teachers misinterpret the meaning of a *p*-value (LessWrongers too); Cohen on this persistent illusion:

What's wrong with NHST? Well, among other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is, "Given these data, what is the probability that *H0* is true?" But as most of us know, what it tells us is "Given that *H0* is true, what is the probability of these (or more extreme) data?" These are not the same…

> (This misunderstanding is incredibly widespread; once you understand it, you'll see it everywhere. I can't count how many times I have seen a comment or blog explaining that a *p*=0.05 means "the probability of the null hypothesis not being true is 95%", in many different variants.)

1. cargo-culting the use of 0.05 as an accept/reject threshold based on historical accident & custom (rather than using a loss function chosen through decision theory to set the threshold based on the cost of false positives).

Similarly, the cargo-culting encourages misuse of two-tailed tests, avoidance of multiple correction, data dredging, and in general, "*p*-value hacking".

1. failing to compare many possible hypotheses or models, and limiting themselves to one - sometimes ill-chosen or absurd - null hypothesis and one alternative
2. deprecating the value of exploratory data analysis and depicting data graphically (see, for example, Anscombe's quartet)
3. ignoring the more important summary statistic of "effect size"
4. ignoring the more important summary statistic of confidence intervals; this is related to how use of *p*-values leads to ignorance of the statistical power of a study - a small study may have only a small chance of detecting an effect if it exists, but turn in misleadingly good-looking *p*-values
5. because null hypothesis tests cannot accept the alternative, but only reject a null, they inevitably cause false alarms upon repeated testing

(An example from my personal experience of the cost of ignoring effect size and confidence intervals: *p*-values cannot (easily) be used to compile a meta-analysis (pooling of multiple studies); hence, studies often do not include the necessary information about means, standard deviations, or effect sizes & confidence intervals which one could use directly. So authors must be contacted, and they may refuse to provide the information or they may no longer be available; both have happened to me in trying to do my dual n-back & iodine meta-analyses.)

Critics' explanations for why a flawed paradigm is still so popular focus on the ease of use and its weakness; from Gigerenzer 2004:

> Hays (1963) had a chapter on Bayesian statistics in the second edition of his widely read textbook but dropped it in the subsequent editions. As he explained to one of us (GG) he dropped the chapter upon pressure from his publisher to produce a statistical cookbook that did not hint at the existence of alternative tools for statistical inference. Furthermore, he believed that many researchers are not interested in statistical thinking in the first place but solely in getting their papers published (Gigerenzer, 2000)…When Loftus (1993) became the editor of *Memory & Cognition*, he made it clear in his editorial that he did not want authors to submit papers in which *p*-, *t*-, or *F*-values are mindlessly being calculated and reported. Rather, he asked researchers to keep it simple and report figures with error bars, following the proverb that "a picture is worth more than a thousand p-values." We admire Loftus for having had the courage to take this step. Years after, one of us (GG) asked Loftus about the success of his crusade against thoughtless significance testing. Loftus bitterly complained that most researchers actually refused the opportunity to escape the ritual. Even when he asked in his editorial letter to get rid of dozens of *p*-values, the authors insisted on keeping them in. There is something deeply engrained in the minds of many researchers that makes them repeat the same action over and over again.

Shifts away from NHST have happened in some fields. Medical testing seems to have made such a shift (I suspect due to the rise of meta-analysis):

> Fidler et al. (2004b, 626) explain the spread of the reform in part by a shift from testing to estimation that was facilitated by the medical literature, unlike psychology, using a common measurement scale, to "strictly enforced editorial policy, virtually simultaneous reforms in a number of leading journals, and the timely re-writing [of] textbooks to fit with policy recommendations." But their description of the process suggests that an accidental factor, the coincidence of several strong-willed editors, also mattered. For the classic collection of papers criticizing significance tests in psychology see Morrison and Hankel (1970) [*The Significance Test Controversy: A Reader*], and for a more recent collection of papers see Harlow et al. (1997) [*What If There Were No Significance Tests?*]. Nickerson (2000) provides a comprehensive survey of this literature.

0.1 Further reading

More on these topics:

- Cohen, "The Earth Is Round (p<.05)" (recommended)
- Effect size FAQ (published as *The Essential Guide to Effect Sizes*, Ellis)
- "The Higgs Boson at 5 Sigmas"
- *The Cult of Statistical Significance*, McCloskey & Ziliak 2008; criticism, their reply

- “Bayesian estimation supersedes the t test”, Kruschke 2012 (see also *Doing Bayesian Data Analysis*); an exposition of a Bayesian paradigm, simulation of false alarm performance compared to his Bayesian code; an excerpt:

The perils of NHST, and the merits of Bayesian data analysis, have been expounded with increasing force in recent years (e.g., W. Edwards, Lindman, & Savage, 1963; Kruschke, 2010b, 2010a, 2011c; Lee & Wagenmakers, 2005; Wagenmakers, 2007).

- A useful bibliography from "A peculiar prevalence of p values just below .05", Masicampo & Lalande 2012:

Although the primary emphasis in psychology is to publish results on the basis of NHST (Cumming et al., 2007; Rosenthal, 1979), the use of NHST has long been controversial. Numerous researchers have argued that reliance on NHST is counterproductive, due in large part because *p* values fail to convey such useful information as effect size and likelihood of replication (Clark, 1963; Cumming, 2008; Killeen, 2005; Kline, 2009 [*Becoming a behavioral science researcher: A guide to producing research that matters*]; Rozeboom, 1960). Indeed, some have argued that NHST has severely impeded scientific progress (Cohen, 1994; Schmidt, 1996) and has confused interpretations of clinical trials (Cicchetti et al., 2011; Ocana & Tannock, 2011). Some researchers have stated that it is important to use multiple, converging tests alongside NHST, including effect sizes and confidence intervals (Hubbard & Lindsay, 2008; Schmidt, 1996). Others still have called for NHST to be completely abandoned (e.g., Carver, 1978).

- `[http://www.gwern.net/DNB%20FAQ#flaws-in-mainstream-science-and-psychology](http://www.gwern.net/DNB%20FAQ#flaws-in-mainstream-science-and-psychology)`

- `[https://www.reddit.com/r/DecisionTheory/](https://www.reddit.com/r/DecisionTheory/)`

Lifeism in the midst of death

> tl;dr: My grandpa died, and I gave a eulogy with a mildly anti-deathist message, in a Catholic funeral service that was mostly pretty disagreeable.

I'm a little uncomfortable writing this post, because it's very personal, and I'm not exactly a regular with friends here. But I need to get it out, and I don't know any other place to put it.

My grandfather (one of two) died last week, and there was a funeral mass (Catholic) today. Although a ‘pro-life’ organisation, the Roman Catholic Church has a very deathist funeral liturgy. It wasn't just ‘Stanley has gone on to a better place’, and all that; the priest had the gall to say that Grandpa had probably done everything that he wanted to do in life, so it was OK for him to die now. I know from discussions with my mother and my aunt that Grandpa did *not* want to die now; although his life and health were not what they used to be, he was happy to live. Yes, he had gone to his great-granddaughter's second birthday party, but he wanted to go to her third, and that will never happen.

There are four of us grandchildren, two (not including me) with spouses. At first, it was suggested that each of us six say one of the Prayers of the Faithful (which are flexible). Mom thought that I might find one that I was willing to recite, so I looked them up online. It wasn't so bad that they end with ‘We pray to the Lord.’ recited by the congregation; I would normally remain silent during that, but I decided that I could say it, and even lead others in saying it, pro forma. And I could endorse the content of some (at least #6 from that list) with some moderate edits. But overall, the whole thing was very disturbing to me. (I had to read HPMoR 45 afterwards to get rid of the bad taste.) I told Mom ‘This is a part of the Mass where I would normally remain in respectful silence.’, and she apologised for ‘put[ting] [me] in an uncomfortable position’ (to quote from our text messages). In the end, the two grandchildren-in-law were assigned to say these prayers.

But we grandchildren still had a place in the programme; we would give eulogies. So I had to think about what to say. I was never close to Grandpa; I loved him well enough, but we didn't have much in common. I tried to think about what I remembered about him and what I would want to tell people about him. It was a little overwhelming; in the end, I read my sibling's notes and decided to discuss only what she did not plan to discuss, and that narrowed it down enough. So then I knew what I wanted to say about Grandpa.

But I wanted to say something more. I wanted to say something to counter the idea that Grandpa's death was OK. I didn't yet know how appalling the priest's sermon would be, but I knew that there would be a lot of excuses made for death. I wanted to preach ‘Grandpa should not have died.’ and go on from there, but I knew that this would be disturbing to people who wanted comfort from their grief, and a lecture on death would not really be a eulogy. Still, I wanted to say something.

(I also didn't want to say anything that could be interpreted as critical of the decision to remove life support. I wasn't consulted on that decision, but under the circumstances, I agree with it. As far as I'm concerned, he was killed on Monday, even though he didn't finally die until Wednesday. In the same conversation in which Mom and I talked about how Grandpa wanted to live, we talked about how he didn't want to live under the circumstances under which he was living on Tuesday, conditions which his doctors expected would never improve. Pulling the plug was the best option available in a bad situation.)

Enough background; here is my eulogy. Some of this is paraphrase, since my written notes were only an outline.

> When I was young, we would visit my grandparents every year, for Thanksgiving or Christmas. Grandma and Grandpa would greet us at the door with hugs and kisses. The first thing that I remember about their place was the candy. Although I didn't realise it at the time, they didn't eat it; it was there as a gift for us kids.
>
> Later I noticed the books that they had, on all topics: religion, history, humour, science fiction, technical material. Most of it was older than I was used to reading, and I found it fascinating. All of this was open to me, and sometimes I would ask Grandpa about some of it; but mostly I just read his books, and to a large extent, this was his influence on me.
>
> Grandpa was a chemical engineer, although he was retired by the time I was able to appreciate that, and this explains the technical material, and to some extent the science fiction. Even that science fiction mostly took death for granted; but Grandpa was with us as long as he was because of the chemists and other people who studied medicine and the arts of healing. They helped him to stay healthy and happy until the heart attack that ended his life.
>
> So, I thank them for what they did for Grandpa, and I wish them success in their future work, to help other people live longer and better, until we never have to go through this again.

I was working on this until the ceremony began, and I even edited it a little in the pew. I wasn't sure until I got up to the podium how strong to make the ending. Ultimately, I said something that could be interpreted as a reference to the Second Coming, but Catholics are not big on that, and my family knows that I don't believe in it. So I don't know how the church officials and Grandpa's personal friends interpreted it, but it could only mean transhumanism to my family.

Nobody said anything, positive or negative, afterwards. Well, a couple of people said that my eulogy was well done; but without specifics, it sounded like they were just trying to make me feel good, to comfort my grief. After my speech, the other three grandchildren went, and then the priest said more pleasant falsehoods, and then it was over.

Goodbye, Grandpa. I wish that you were alive and happy in Heaven, but at least you were alive and happy here on Earth for a while. I'll miss you.

[Edit: Fix my cousin's age.]

My workflow

Over the last 6 months I've started doing a lot of things differently. Some of these changes seem to have increased my work output a good bit and made me happier. I normally hesitate to share habits, but I'm pretty happy with these in particular, and even if they will work for only a few people I think they are worth sharing. Most of the habits I've adopted are fairly common, but I hope I can help people anyway by identifying the habits that have most helped me.

I'm curious to hear about alternatives that have worked for you.

Workflowy:

Workflowy lets you edit a single collapsible outline. I use it very extensively. It is much more convenient than the network of google docs it replaced, and I use it much more often. It is much like other outliners, but (1) has a slicker interface, (2) works offline, (3) lets you recurse on and share sublists.

Workflowy is free to try but costs $5 a month. This may seem expensive for what it does, but if you use (or could use!) outliners a lot this is not enough to matter. After some searching Workflowy seems like the best option. I'm sure I like Workflowy more than most people, but I *really* like it, so I think it's worth trying.

Here is a skeleton of my workflowy list, which hosts many of the other systems in this post.

Checklists:

I have a checklist of tasks to do each night before sleeping. In the past I would often forget one of these things; putting them in a checklist helps me do them more reliably and makes me more relaxed.

Checklists for other occasions, particularly waking up and traveling, are also helpful, but are much less important to me.

Todo lists:

I now maintain two todo lists: one with a list of tasks for each upcoming day, and one with a list of tasks for future events ("I'm in the UK," "it is Thursday," "I'm going grocery shopping"). Whenever I think of something I should do, I either put it under a future day and do it when that day arrives, or I put it with an associated event. Each night I check both lists and decide what to do tomorrow.

Beeminder:

Beeminder is a service that holds you to commitments and tracks your progress. It has helped me a lot over the last months. I've experimented with a few different commitments, but two have been most useful: following a daily routine, and doing a minimum amount of work each day (on average). Beeminder has pretty low overhead.

Reflection:

I spend about 10% of my productive time reflecting on how things have been going and what I should do differently. I benefit from producing concrete possible changes each time I sit down to think. I realized how important this is for me recently; since I've started doing it more reliably, I have gotten a lot more out of reflection.

Pomodoro:

I do my work in uninterrupted blocks of 20 minutes, punctuated by 2-3 minute breaks. This is my bastardized, minimalist version of the pomodoro technique, which I arrived at by trial and error. I use Alinof timer, which was recommended to me by a friend.

Calendar:

I now record commitments on my calendar reliably and check it each night. I failed to do this for 6 months after finishing my undergraduate degree, which I think was a serious mistake. I became much more reliable at checking my calendar after adopting a daily checklist.

Time Logging:

Whenever I start a new activity, I write down the current time and a description of what I just stopped doing. At the end of the day I spend a few minutes reading this log and estimating how much time I spent on each activity. This makes me more attentive to time during the day, helps me remember what I did throughout the day, and frees up attention. Sometimes I use the logs to try and notice trends. For example, I've been exercising on random days and measuring how this affects my time. I don't yet know if this helps at all.

Catch:

Catch is a note-taking app. It is very minimal, and lets you record a voice note by pressing a single button. It has substantially increased my affordance for taking notes during the day, which I use to remember todo items and help with time logging.

Three kinds of moral uncertainty

Related to: Moral uncertainty (wiki), Moral uncertainty - towards a solution?, Ontological Crisis in Humans.

> **Moral uncertainty** (or **normative uncertainty**) is uncertainty about how to act given the diversity of moral doctrines. For example, suppose that we knew for certain that a new technology would enable more humans to live on another planet with slightly less well-being than on Earth[1]. An average utilitarian would consider these consequences bad, while a total utilitarian would endorse such technology. If we are uncertain about which of these two theories are right, what should we do? (LW wiki)

I have long been slightly frustrated by the existing discussions about moral uncertainty that I've seen. I suspect that the reason has been that they've been unclear on what exactly they *mean* when they say that we are "uncertain about which theory is right" - what *is* uncertainty about moral theories? Furthermore, especially when discussing things in an FAI context, it feels like several different senses of moral uncertainty get mixed together. Here is my suggested breakdown, with some elaboration:

Descriptive moral uncertainty. *What is the most accurate way of describing my values?* The classical FAI-relevant question, this is in a sense the most straightforward one. We have some set values, and although we can describe parts of them verbally, we do not have conscious access to the deep-level cognitive machinery that generates them. We might feel relatively sure that our moral intuitions are produced by a system that's mostly consequentialist, but suspect that parts of us might be better described as deontologist. A solution to descriptive moral uncertainty would involve a system capable of somehow extracting the mental machinery that produced our values, or creating a moral reasoning system which managed to produce the same values by some other process.

Epistemic moral uncertainty. *Would I reconsider any of my values if I knew more?* Perhaps we hate the practice of eating five-sided fruit and think that everyone who eats five-sided fruit should be thrown to jail, but if we found out that five-sided fruit made people happier and had no averse effects, we would change our minds. This roughly corresponds to the "our wish if we knew more, thought faster" part of Eliezer's original CEV description. A solution to epistemic moral uncertainty would involve finding out more about the world.

Intrinsic moral uncertainty. *Which axioms should I endorse?* We might be intrinsically conflicted between different value systems. Perhaps we are trying to choose whether to be loyal to a friend or whether to act for the common good (a conflict between two forms of deontology, or between deontology and consequentialism), or we could be conflicted between positive and negative utilitarianism. In its purest form, this sense of moral uncertainty closely resembles what would otherwise be called a wrong question, one where

> you cannot even *imagine* any concrete, specific state of how-the-world-is that would answer the question. When it doesn't even seem *possible* to answer the question.

But unlike wrong questions, questions of intrinsic moral uncertainty are real ones that you need to actually answer in order to make a choice. They are generated when different modules within your brain generate different moral intuitions, and are essentially power struggles between various parts of your mind. A solution to intrinsic moral uncertainty would involve somehow tipping the balance of power in favor of one of the "mind factions". This could involve developing an argument sufficiently persuasive to convince most parts of yourself, or self-modifying in such a way that one of the factions loses its sway over your decision-making. (Of course, if you already knew for certain which faction you wanted to expunge, you wouldn't need to do it in the first place.) I would roughly interpret the "our wish ... if we had grown up farther together" part of CEV to be an attempt to model some of the social influences on our moral intuitions and thereby help resolve cases of intrinsic moral uncertainty.

This is a very preliminary categorization, and I'm sure that it could be improved upon. There also seem to exist cases of moral uncertainty which are hybrids of several categories - for example, ontological crises seem to be mostly about intrinsic moral uncertainty, but to also incorporate some elements of epistemic moral uncertainty. I also have a general suspicion that these categories still don't cut reality that well at the joints, so any suggestions for improvement would be much appreciated.

Ritual 2012: A Moment of Darkness

This is the second post of the 2012 Ritual Sequence. The Introduction post is here.

This is... the extended version, I suppose, of a speech I gave at the Solstice.

The NYC Solstice Weekprior celebration begins bright and loud, and gradually becomes somber and poignant. Our opening songs are about the end of the world, but in a funny, boisterous manner that gets people excited and ready to sing. We gradually wind down, dimming lights, extinguishing flames. We turn to songs that aren't sad but are more quiet and pretty.

And then things get grim. We read Beyond the Reach of God. We sing songs about a world where we are alone, where there is nothing protecting us, and where we somehow need to survive and thrive, even when it looks like the light is failing.

We extinguish all but a single candle, and read an abridged version of the Gift We Give to Tomorrow, which ends like this:

Once upon a time,
far away and long ago,
there were intelligent beings who were not themselves intelligently designed.

Once upon a time,
there were lovers, created by something that did not love.

Once upon a time,
when all of civilization was a single galaxy,

A single star.
A single planet.
A place called Earth.

Once upon a time.

And then we extinguish that candle, and sit for a moment in the darkness.

This year, I took that time to tell a story.

It's included in the 2012 Ritual Book. I was going to post it at the end of the sequence. But I realized that it's actually pretty important to the "What Exactly is the Point of Ritual?" discussion. So I'm writing a more fleshed out version now, both for easy reference and for people who don't feel like hunting through a large pdf to find it.

It's a bit longer, in this version - it's what I might have said, if time wasn't a constraint during the ceremony.

A year ago, I started planning for tonight. In particular, for this moment, after the last candle is snuffed out and we're left alone in the dark with the knowledge that our world is unfair and that we have nobody to help us but each other.

I wanted to talk about death.

My grandmother died two years ago. The years leading up to her death were painful. She slowly lost her mobility, until all she could do was sit in her living room and hope her family would come by to visit and talk to her.

Then she started losing her memory, so she had a hard time even having conversations at all. We tried to humor her, but there's only so many times you can repeat the same thought in a five minute interval before your patience wears thin, and it shows, no matter how hard you try.

She lost her rationality, regressing into a child who would argue petulantly with my mother about what to eat, and when to exercise, and visit her friends. She was a nutritionist, she knew what she was supposed to eat and why. She knew how to be healthy. And she wanted to be healthy. But she lost her ability to negotiate her near term and long term desires on her own.

Eventually even deciding to eat at all became painful. Eventually even forming words became exhausting.

Eventually she lost not just her rationality, but her agency. She stopped making decisions. She lay on her bed in the hospital, not even having the strength to complain anymore. My mother got so excited on days when she argued petulantly because at least she was doing *something*.

She lost everything that I thought made a person a person, and I stopped thinking of her as one.

Towards the end of her life, I was visiting her at the hospital. I was sitting next to her, being a dutiful grandson. Holding her hand because I knew she liked that. But she seemed like she was asleep, and after 10 minutes or so I got bored and said "alright, I'm going to go find Mom now. I'll be back soon."

And she squeezed my hand, and said "No, stay."

Those two words were one of the last decisions she ever made. One of the last times she had a desire about how her future should be. She made an exhausting effort to turn those desires into words and then breath those words into sounds so that her grandson would spend a little more time with her.

And I was so humiliated that I had stopped believing that inside of this broken body and broken mind was a person who still desperately wanted to be loved.

She died a week or two later.

Her funeral was a Catholic Mass. My mom had made me go to Mass as a child. It always annoyed me. But in that moment, I was so grateful to be able to hold hands with a hundred people, for all of us to speak in unison, without having to think about it, and say:

"Our father, who art in heaven, hallowed by thy name. Thy kingdom come, thy will be done, on earth as it is in heaven. Give us this day our daily bread, and forgive us of our trespasses, as we forgive those who trespass against us. And lead us not into temptation, but deliver us from evil."

I'm not sure if having that one moment of comforting unity was worth 10 years of attending Catholic mass.

It's a legitimately hard question. I don't know the answer.

But I was still so frustrated that this comforting ritual was all based on falsehoods. There's plenty of material out there you can use to create a beautiful secular funeral, but it's not just about having pretty or powerful words to say. It's about about knowing the words already, having them already be part of you and your culture and your community.

Because when somebody dies, you don't have time or energy for novelty. You don't want to deal with new ideas that will grate slightly against you just because they're new. You want cached wisdom that is simple and beautiful and true, that you share with others, so that when something as awful as death happens to you, you have tools to face it, and you don't have to face it alone.

I was thinking about all that, as I prepared for this moment.

But my Grandmother's death was a long time ago. I wanted the opportunity to process it in my own way, in a community that shared my values. But it wasn't really a pressing issue that bore down on me. Dealing with death felt important, but it was a sort of abstract importance.

And then, the second half of this year happened.

A few months ago, an aspiring rationalist friend of mine e-mailed me to tell me that a relative died. They described the experience of the funeral, ways in which it was surprisingly straightforward, and other ways in which it was very intense. My friend had always considered themselves an anti-deathist, but it was suddenly very real to them. And it sort of sank in for me too - death is still a part of this world, and our community doesn't really have ways to deal with it.

And then, while I was still in the middle of the conversation with that friend, I learned that another friend had lost somebody, that same day.

Later, I would learn that a coworker of mine also lost somebody that day as well.

Death was no longer abstract. It was real, painfully real, even if I myself didn't know the people who died. My friends were hurting, and I felt their pain.

I wandered off into the night to sing my Stonehenge song by myself. It's not quite good enough at what I needed it for - I'm not a skilled enough songwriter to write that song, yet. But it's the only song I know of that attempts to do what I needed. To grimly acknowledge this specific adversary, to not offer any false hope about the inevitability of our victory, but to nonetheless march onward, bitterly determined that not quite so many people will die tomorrow as today.

I came back inside. I chatted with another friend about the experience. She offered me what comfort she could. She attempted to offer some words to the effect of “well, death has a purpose sometimes. It helps you see the good things -”

Gah, I thought.

What’s interesting is that I’m not actually that much of an anti-deathist. I think our community’s obsession with eliminating death without regard for the consequences is potentially harmful. I think there are, quite frankly, worse things in the world. If I had to choose between my Grandmother not dying, and my Grandmother not having to gradually lose everything she thought made her *her* until her own grandson forgot that she was a person, spending her days wracked with pain, I would probably choose the latter.

But still, I’ve come to accept that death is bad, unequivocally bad, even if some things are worse. And I had sort of forgotten, since I’m often at odds with other Less Wrongers about this, how big the gulf was between us and the rest of the world.

I didn’t hold it against my friend. She meant well, and having someone to talked to helped.

A week later, a friend of hers died.

A week after that, another friend of mine lost somebody.

A week after that, it wasn’t a direct friend of a friend who died, but a local activist was murdered a few blocks from someone’s house, and they cancelled plans with me because they were so upset.

Then a hurricane hit New York. Half the city went dark. While it was unrelated, at least one of my friends experienced a death, of sorts, that week. And even if none of my friends were directly hurt by Hurricane Sandy, you couldn’t escape the knowledge that there were people who weren’t so lucky.

And I went back to my notes I had written for this moment and stared and them and thought...

…

…fuck.

Winter was coming and I didn’t know what to do. Death is coming, and our community isn’t ready. I set out to create a holiday about death and... it turns out that’s a lot of responsibility, actually.

This was important, this was incredibly important and so incredibly hard to handle correctly. We as a community - the New York community, at least - need a way to process what happened to us this year, but what happened to each of us is personal and even though most of share the same values we all deal with death in our own way and… and... and somehow after all of that, after taking a moment to process it, we need to climb back out of that darkness and end the evening feeling joyful and triumphant and proud to be human, without resorting to lies.

…

…

…there's a lot I don't know how yet, about what to do, or what to say.

But here's what I do know:

My grandmother died. But she lived to her late eighties. She had a family of 5 children who loved her. She had a life full of not just fun and travel and adventure but of scientific discovery. She was a dietitian. She helped do research on diabetes. She was an inspiration to women at a time when a woman being a researcher was weird and a big deal. When I say she had a long, full life, I'm not just saying something nice sounding.

My grandmother won at life, by any reasonable standard.

Not everyone gets to have that, but my grandmother did. She was the matriarch of a huge extended family that all came home for Christmas eve each year, and sang songs and shared food and loved each other. She died a few weeks after Christmas, and that year, everyone came to visit, and honestly it was one of the best experiences of my life.

In the dead of winter, each year, two dozen of people came to Poughkeepsie, to a big house sheltered by a giant cottonwood tree, and were able to celebrate *without* worrying about running out of food in the spring. At the darkest time of the year, my mother ran lights up a hundred foot tall pine tree that you could see for miles.

We were able to eat because hundreds of miles away, mechanical plows tilled fields in different climates, producing so much food that we literally could feed the entire world if we could solve some infrastructure and economic problems.

We were able to drive to my grandmother's house because other mechanical plows crawled through the streets all night, clearing the ice and snow away.

Some of us were able to come to my grandmothers house from a thousand miles away, flying through the sky, higher than ancient humans even imagined angels might live.

And my Grandmother died in her late eighties, but she also *didn't* die when she was in her 70s and the cancer first struck her. Because we had chemotherapy, and host of other tools to deal with it.

And the most miraculous amazing thing is that this isn't a miracle. This isn't a mystery. We know how it came to be, and we have the power to learn to understand it even better, and do more.

In this room, right now, are people who take this all seriously. Dead seriously, who don't just shout "Hurrah humanity" because shouting things together in a group is fun.

We have people in this room, right now, who are working on fixing big problems in the medical industry. We have people in this room who are trying to understand and help fix the criminal justice system. We have people in this room who are dedicating their lives to eradicating global poverty. We have people in this room who are literally working to set in motion plans to optimize *everything ever*. We have people in this room who are working to make sure that the human race doesn't destroy itself before we have a chance to become the people we really want to be.

And while they aren't in this room, there are people we know who would be here if they could, who are doing their part to try and solve this whole death problem once and for all.

And I don't know whether and how well any of us are going to succeed at any of these things, but...

God damn, people. You people are amazing, and even if only one of you made a dent in some of the problems you're working on, that… that would just be incredible.

And there are people in this room who aren't working on anything that grandiose. People who aren't trying to solve death or save the world from annihilation or alleviate suffering on a societal level. But who spend their lives making art. Music. Writing things sometimes.

People who fill their world with beauty and joy and enthusiasm, and pies and hugs and games and… and I don't have time to give a shout out to everyone in the room but you all know who you are.

This room is full of people who spend their lives making this world less ugly, less a sea of blood and violence and mindless replication. People who are working to make tomorrow brighter than today, in one way or another.

And I am so proud to know all of you, to have you be a part of my life, and to be a part of yours.

I love you.

You make this world the sort of place I'd want to keep living, forever, if I could.

The sort of world I'd want to take to the stars.

By Which It May Be Judged

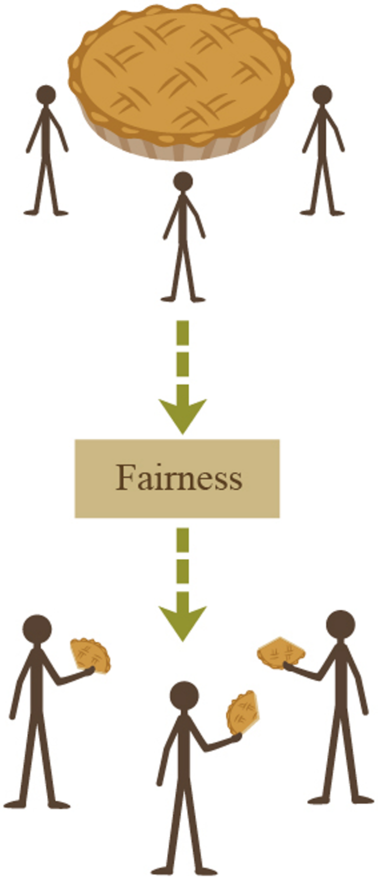
Followup to: Mixed Reference: The Great Reductionist Project

> 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

Suppose three people find a pie - that is, three people exactly simultaneously spot a pie which has been exogenously generated in unclaimed territory. Zaire wants the entire pie; Yancy thinks that 1/3 each is fair; and Xannon thinks that fair would be taking into equal account everyone's ideas about what is "fair".

I myself would say unhesitatingly that a third of the pie each, is fair. "Fairness", as an ethical concept, can get a lot more complicated in more elaborate contexts. But in this simple context, a lot of other things that "fairness" could depend on, like work inputs, have been eliminated or made constant. Assuming no relevant conditions other than those already stated, "fairness" simplifies to the mathematical procedure of splitting the pie into equal parts; and when this logical function is run over physical reality, it outputs "1/3 for Zaire, 1/3 for Yancy, 1/3 for Xannon".

Or to put it another way - just like we get "If Oswald hadn't shot Kennedy, nobody else would've" by running a logical function over a true causal model - similarly, we can get the hypothetical 'fair' situation, whether or not it actually happens, by running the physical starting scenario through a logical function that describes what a 'fair' outcome would look like:



So am I (as Zaire would claim) just assuming-by-authority that I get to have everything my way, since I'm not defining 'fairness' the way *Zaire* wants to define it?

No more than mathematicians are flatly ordering everyone to assume-without-proof that two different numbers can't have the same successor. For fairness to be what everyone thinks is "fair" would be *entirely* circular, structurally isomorphic to "Fzeem is what everyone thinks is fzeem"... or like trying to define the counting numbers as "whatever anyone thinks is a number". It only even *looks* coherent because everyone secretly already has a mental picture of "numbers" - because their brain already navigated to the referent. But *something* akin to axioms is needed to talk about "numbers, as opposed to something else" in the first place. Even an inchoate mental image of "0, 1, 2, ..." implies the axioms no less than a formal statement - we can extract the axioms back out by asking questions about this rough mental image.

Similarly, the intuition that fairness has *something* to do with dividing up the pie equally, plays a role akin to secretly already having "0, 1, 2, ..." in mind as the subject of mathematical conversation. You need axioms, not as assumptions that aren't justified, but as pointers to what the heck the conversation is supposed to be *about.*

Multiple philosophers have suggested that this stance seems similar to "rigid designation", i.e., when I say 'fair' it intrinsically, rigidly refers to something-to-do-with-equal-division. I confess I don't see it that way myself - if somebody thinks of Euclidean geometry when you utter the sound "num-berz" they're not doing anything false, they're associating the sound to a different logical thingy. It's not about words with intrinsically rigid referential power, it's that the words are *window dressing* on the underlying entities. I want to *talk about* a particular *logical entity,* as it might be defined by either axioms or inchoate images, regardless of which word-sounds may be associated to it. If you want to call that "rigid designation", that seems to me like adding a level of indirection; I don't care about the *word* 'fair' in the first place, I care about the logical entity of fairness. (Or to put it even more sharply: since my ontology does not have room for physics, logic, *plus* designation, I'm not very interested in discussing this 'rigid designation' business unless it's being reduced to something else.)

Once issues of justice become more complicated and all the contextual variables get added back in, we might not be sure if a *disagreement* about 'fairness' reflects:

1. The equivalent of a multiplication error within the same axioms - incorrectly dividing by 3. (Or more complicatedly: You might have a sophisticated axiomatic concept of 'equity', and *incorrectly* process those axioms to invalidly yield the assertion that, in a context where 2 of the 3 must starve and there's only enough pie for at most 1 person to survive, you should still divide the pie equally instead of flipping a 3-sided coin. Where I'm assuming that this conclusion is 'incorrect', not because I disagree with it, but because it didn't actually follow from the axioms.)
2. Mistaken models of the physical world fed into the function - mistakenly thinking there's 2 pies, or mistakenly thinking that Zaire has no subjective experiences and is not an object of ethical value.
3. People associating different logical functions to the letters F-A-I-R, which isn't a *disagreement about* some common pinpointed variable, but just different people wanting different things.

There's a lot of people who feel that this picture leaves out something fundamental, especially once we make the jump from "fair" to the broader concept of "moral", "good", or "right". And it's this worry about leaving-out-something-fundamental that I hope to address next...

...but please note, if we confess that 'right' lives in a world of physics and logic - because *everything* lives in a world of physics and logic - then we *have* to translate 'right' into those terms *somehow.*

And that is the answer Susan should have given - if she could talk about sufficiently advanced epistemology, sufficiently fast - to Death's entire statement:

> 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. And yet — Death waved a hand. And yet you act as if there is some ideal order in the world, as if there is some ... *rightness* in the universe by which it may be judged.

"But!" Susan should've said. "When we judge the universe we're comparing it to a *logical* referent, a sort of thing that isn't *in* the universe! Why, it's just like looking at a heap of 2 apples and a heap of 3 apples on a table, and comparing their invisible product to the number 6 - there isn't any 6 if you grind up the whole table, even if you grind up the whole universe, but the product is *still* 6, physico-logically speaking."

If you require that Rightness be written on some particular great Stone Tablet somewhere - to be "a light that shines from the sky", outside people, as a different

Terry Pratchett book put it - then indeed, there's no such Stone Tablet anywhere in our universe.

But there *shouldn't* be such a Stone Tablet, *given* standard intuitions about morality. This follows from the Euthryphro Dilemma out of ancient Greece.

The original Euthryphro dilemma goes, "Is it pious because it is loved by the gods, or loved by the gods because it is pious?" The religious version goes, "Is it good because it is commanded by God, or does God command it because it is good?"

The standard atheist reply is: "Would you say that it's an intrinsically good thing - even if the event has no further causal consequences which are good - to slaughter babies or torture people, if that's what God says to do?"

If we can't make it good to slaughter babies by tweaking the state of God, then morality doesn't come from God; so goes the standard atheist argument.

But if you can't make it good to slaughter babies by tweaking the physical state of *anything* - if we can't imagine a world where some great Stone Tablet of Morality has been physically rewritten, and what is right has changed - then this is telling us that...

(drumroll)

...what's "right" is a logical thingy rather than a physical thingy, that's all. The mark of a logical validity is that we can't concretely visualize a coherent possible world where the proposition is false.

And I mention this in hopes that I can show that it is not moral anti-realism to say that moral statements take their truth-value from logical entities. Even in Ancient Greece, philosophers implicitly knew that 'morality' ought to be such an entity - that it *couldn't* be something you found when you ground the Universe to powder, because then you could resprinkle the powder and make it wonderful to kill babies - though they didn't know how to say what they knew.

There's a lot of people who still feel that Death *would* be right, if the universe were all physical; that the kind of dry logical entity I'm describing here, isn't sufficient to carry the bright alive feeling of goodness.

And there are others who accept that physics and logic is everything, but who - I think *mistakenly* - go ahead and also accept Death's stance that this makes morality a lie, or, in lesser form, that the bright alive feeling can't make it. (Sort of like people who accept an incompatibilist theory of free will, also accept physics, and conclude with sorrow that they are indeed being controlled by physics.)

In case anyone is bored that I'm *still* trying to fight this battle, well, here's a quote from a recent Facebook conversation with a famous early transhumanist:

> No doubt a "crippled" AI that didn't understand the existence or nature of first-person facts could be nonfriendly towards sentient beings... Only a zombie wouldn't value Heaven over Hell. For reasons we simply don't understand, the negative value and normative aspect of agony and despair is built into the nature of the experience itself. Non-reductionist? Yes, on a standard materialist ontology. But not IMO within a more defensible Strawsonian physicalism.

It would actually be *quite surprisingly helpful* for increasing the percentage of people who will participate meaningfully in saving the planet, if there were some reliably-working standard explanation for why physics and logic together have enough room to contain morality. People who think that reductionism means we have to lie to our children, as Pratchett's Death advocates, won't be much enthused about the Center for Applied Rationality. And there are a fair number of people out there who still advocate proceeding in the confidence of ineffable morality to construct sloppily designed AIs.

So far I don't know of any exposition that works reliably - for the thesis for how morality *including* our intuitions about whether things *really are justified* and so on, is preserved in the analysis to physics plus logic; that morality has been explained rather than explained away. Nonetheless I shall now take another stab at it, starting with a simpler bright feeling:

When I see an unusually neat mathematical proof, unexpectedly short or surprisingly general, my brain gets a joyous sense of *elegance.*

There's presumably some functional slice through my brain that implements this emotion - some configuration subspace of spiking neural circuitry which corresponds to my *feeling* of elegance. Perhaps I should say that elegance is *merely* about my brain switching on its elegance-signal? But there are concepts like Kolmogorov complexity that give more formal meanings of "simple" than "Simple is whatever makes my brain feel the emotion of simplicity." Anything you do to fool my brain wouldn't make the proof *really* elegant, not in that sense. The emotion is not free of semantic content; we could build a correspondence theory for it and navigate to its logical+physical referent, and say: "Sarah feels like this proof is elegant, and her feeling is *true.*" You could even say that certain proofs are elegant even if no conscious agent sees them.

My description of 'elegance' admittedly did invoke agent-dependent concepts like 'unexpectedly' short or 'surprisingly' general. It's almost certainly true that with a different mathematical background, I would have different standards of elegance and experience that feeling on *somewhat* different occasions. Even so, that still seems like moving around in a field of *similar* referents for the emotion - much more similar to each other than to, say, the distant cluster of 'anger'.

Rewiring my brain so that the 'elegance' sensation gets activated when I see mathematical proofs where the words have lots of vowels - that wouldn't *change* what is elegant. Rather, it would make the feeling be *about* something else entirely; different semantics with a different truth-condition.

Indeed, it's not clear that this thought experiment is, or should be, *really* conceivable. If all the associated computation is about vowels instead of elegance, then from the inside you would expect that to *feel vowelly*, not *feel elegant*...

...which is to say that even feelings can be associated with logical entities. Though unfortunately not in any way that will *feel like* qualia if you can't read your own source code. I could write out an exact description of your visual cortex's spiking code for 'blue' on paper, and it wouldn't actually *look* blue to you. Still, on the higher level of description, it should seem intuitively plausible that if you tried rewriting the relevant part of your brain to count vowels, the resulting sensation would no longer have the content or even the *feeling* of elegance. It would compute vowelliness, and feel vowelly.

My feeling of mathematical elegance is motivating; it makes me more likely to search for similar such proofs later and go on doing math. You could construct an agent that tried to add more vowels instead, and if the agent asked itself why it was doing that, the resulting justification-thought wouldn't *feel like* because-it's-elegant, it would *feel like* because-it's-vowelly.

In the same sense, when you try to do what's right, you're motivated by things like (to yet again quote Frankena's list of terminal values):

> "Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc."

If we reprogrammed you to count paperclips instead, it wouldn't feel like *different* things having the *same* kind of motivation behind it. It wouldn't feel like doing-what's-right for a different guess about what's right. It would feel like doing-what-leads-to-paperclips.

And I quoted the above list because the feeling of rightness isn't *about* implementing a particular logical function; it contains no mention of logical functions at all; in the environment of evolutionary ancestry nobody has *heard* of axiomatization; these feelings are *about* life, consciousness, etcetera. If I could write out the whole truth-condition of the feeling in a way you could compute, you would still feel Moore's Open Question: "I can see that this event is high-rated by logical function X, but is X really *right?*" - since you can't read your own source code and the description wouldn't be commensurate with your brain's native format.

"But!" you cry. "But, is it really *better* to do what's right, than to maximize paperclips?" Yes! As soon as you start trying to cash out the logical function that gives betterness its truth-value, it will output "life, consciousness, etc. $>_B$ paperclips". And if your brain were computing a different logical function instead, like makes-more-paperclips, it wouldn't feel *better,* it would feel *moreclippy.*

But is it really *justified* to keep our own sense of betterness? Sure, and that's a logical fact - it's the objective output of the logical function corresponding to your experiential sense of what it means for something to be 'justified' in the first place. This doesn't mean that Clippy the Paperclip Maximizer will self-modify to do only things that are justified; Clippy doesn't judge between self-modifications by computing justifications, but rather, computing *clippyflurphs.*

But isn't it *arbitrary* for Clippy to maximize paperclips? Indeed; once you implicitly or explicitly pinpoint the logical function that gives judgments of arbitrariness their truth-value - presumably, revolving around the presence or absence of justifications - then this logical function will objectively yield that there's no justification whatsoever for maximizing paperclips (which is why *I'm* not going to do it) and hence that Clippy's decision is arbitrary. Conversely, Clippy finds that there's no clippyflurph for preserving life, and hence that it is unclipperiffic. But unclipperifficness isn't arbitrariness any more than the number 17 is a right triangle; they're different logical entities pinned down by different axioms, and the corresponding judgments will have different semantic content and *feel different.* If Clippy is architected to experience that-which-you-call-qualia, Clippy's feeling of clippyflurph will be *structurally* different from the way justification feels, not just red versus blue, but vision versus sound.

But surely one *shouldn't* praise the clippyflurphers rather than the just? I quite agree; and as soon as you navigate referentially to the coherent logical entity that is the truth-condition of *should* - a function on potential actions and future states - it will agree with you that it's better to avoid the arbitrary than the unclipperiffic. Unfortunately, this logical fact does not correspond to the truth-condition of any meaningful proposition computed by Clippy in the course of how it efficiently transforms the universe into paperclips, in much the same way that rightness plays no role in that-which-is-maximized by the blind processes of natural selection.

Where moral judgment is concerned, it's logic all the way down. *ALL* the way down. Any frame of reference where you're worried that it's *really* no better to do what's right then to maximize paperclips... well, that *really* part has a truth-condition (or what does the "really" mean?) and as soon as you write out the truth-condition you're going to end up with yet another ordering over actions or algorithms or meta-algorithms or *something.* And since grinding up the universe won't and *shouldn't* yield any miniature '>' tokens, it must be a *logical* ordering. And so whatever logical ordering it is you're worried about, it probably *does* produce 'life > paperclips' - but Clippy isn't computing that logical fact any more than your pocket calculator is computing it.

Logical facts have no power to directly affect the universe except when some part of the universe is computing them, and morality is (and *should* be) logic, not physics.

Which is to say:

> The old wizard was staring at him, a sad look in his eyes. "I suppose I *do* understand now," he said quietly.
>
> "Oh?" said Harry. "Understand what?"
>
> "Voldemort," said the old wizard. "I understand him now at last. Because to believe that the world is truly like that, you must believe there is no justice in it, that it is woven of darkness at its core. I asked you why he became a monster, and you could give no reason. And if I could ask *him*, I suppose, his answer would be: Why not?"
>
> They stood there gazing into each other's eyes, the old wizard in his robes, and the young boy with the lightning-bolt scar on his forehead.
>
> "Tell me, Harry," said the old wizard, "will *you* become a monster?"
>
> "No," said the boy, an iron certainty in his voice.
>
> "Why not?" said the old wizard.
>
> The young boy stood very straight, his chin raised high and proud, and said: "There is no justice in the laws of Nature, Headmaster, no term for fairness in the equations of motion. The universe is neither evil, nor good, it simply does not

care. The stars don't care, or the Sun, or the sky. But they don't have to! *We* care! There *is* light in the world, and it is *us!*"

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

Next post: "Standard and Nonstandard Numbers"

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

So you think you understand Quantum Mechanics

This post is prompted by the multitude of posts and comments here using quantum this and that in an argument (quantum dice, quantum immortality, quantum many worlds...). But how does one know if they understand the concept they use? In school a student would have to write a test and get graded. It strikes me as a reasonable thing to do here, as well: let people test their understanding of the material so that they can calibrate their estimate of their knowledge of the topic. This is an attempt to do just that.

Let's look at one of the very first experiments demonstrating that in the microscopic world things are usually quantized: the Stern-Gerlach experiment, in which measured angular momentum is shown to take discrete values. The gist of the experiment is that in a varying magnetic field the tidal force on a magnet is not perfectly balanced and so the magnet moves toward or away from the denser field, depending on the orientation of its poles. This is intuitively clear to anyone who ever played with magnets: the degree of attraction or repulsion depends on the relative orientation of the magnets (North pole repels North pole etc.). It is less obvious that this effect is due to the spatially varying magnetic field density, but it is nonetheless the case.

In the experiment, one magnet is large (the S-G apparatus itself) and one is small (a silver atom injected into the magnetic field of the large magnet). The experiment shows that an unoriented atom suddenly becomes aligned either along or against the field, but not in any other direction. It's like a compass needle that would only be able to point North and South (and potentially in a few other directions) but not anywhere in between.

If necessary, please read through the more detailed description of the experiment on Wikipedia or in any other source before attempting the following questions (usually called meditations in the idiosyncratic language used on this forum).

Meditation 1. When exactly does the atom align itself? As soon as it enters the field? At some random moment as it travels through the field? The instance it hits the screen behind the field? In other words, in the MWI picture, when does the world split into two, one with the atom aligned and one with the atom anti-aligned? In the Copenhagen picture, does the magnetic field measure the atom spin, and if so, when, or does the screen do it?

Hint. Consider whether/how you would tell these cases apart experimentally.

Meditation 2. Suppose you make two holes in the screen where the atoms used to hit it, then merge the atoms into a single stream again by applying a reverse field. Are the atoms now unaligned again, or 50/50 aligned/anti-aligned or something else?

Hint. What's the difference between these cases?

Meditation 3. Suppose that instead of the reversing field in the above experiment you keep the first screen with two holes in it, and put a second screen (without any holes) somewhere behind the first one. What would you expect to see on the second screen and why? Some possible answers: two equally bright blobs corresponding to aligned and anti-aligned atoms respectively; the interference pattern from each atom passing through both holes at once, like in the double-slit experiment; a narrow single blob in the center of the second screen, as if the atoms did not go through the first part of the apparatus at all; a spread-out blob with a maximum at the center, like you would expect from the classical atoms.

Hint. Consider/reconsider your answer to the first two questions.

Meditation 4. Suppose you want to answer M1 experimentally and use an extremely sensitive accelerometer to see which way each atom is deflecting **before** it hits the screen by measuring the recoil of the apparatus. What would you expect to observe?

Hint. Consider a similar setup for the double-slit experiment.

This test is open-book and there is no time limit. You can consult any sources you like, including textbooks, research papers, your teachers, professional experimental or theoretical physicists, your fellow LWers, or the immortal soul of Niels Bohr through your local medium. If you have access to the Stern-Gerlach apparatus in your physics lab, feel free to perform any experiments you may find helpful. As they say, if you are not cheating, you are not trying hard enough.

By the way, if anyone wants to supply the pictures to make the setup for each question clearer, I'd be more than happy to include them in this post. If anyone wants to turn the meditations into polls, please do so in the comments.

Footnote: not posting this in Main, because I'm not sure how much interest there is here for QM questions like this.

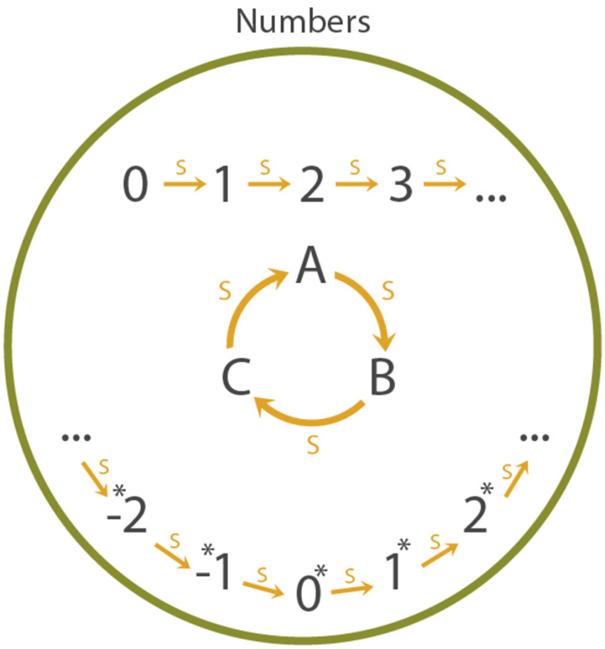
Standard and Nonstandard Numbers

Followup to: Logical Pinpointing

"Oh! Hello. Back again?"

Yes, I've got another question. Earlier you said that you *had* to use second-order logic to define the numbers. But I'm pretty sure I've heard about something called 'first-order Peano arithmetic' which is also supposed to define the natural numbers. Going by the name, I doubt it has any 'second-order' axioms. Honestly, I'm not sure I understand this second-order business at all.

"Well, let's start by examining the following model:"



"This model has three properties that we would expect to be true of the standard numbers - 'Every number has a successor', 'If two numbers have the same successor they are the same number', and '0 is the only number which is not the successor of any number'. All three of these statements are true in this model, so in that sense it's quite numberlike -"

And yet this model clearly is *not* the numbers we are looking for, because it's got all these mysterious extra numbers like C and -2*. That C thing even loops around, which I certainly wouldn't expect any number to do. And then there's that infinite-in-both-directions chain which isn't corrected to anything else.

"Right, so, the difference between first-order logic and second-order logic is this: In first-order logic, we can get rid of the ABC - make a statement which *rules out* any model that has a loop of numbers like that. But we can't get rid of the infinite chain underneath it. In second-order logic we can get rid of the extra chain."

I would ask you to explain why that was true, but at this point I don't even know what second-order logic *is.*

"Bear with me. First, consider that the following formula *detects 2-ness:*"

$$x + 2 = x * 2$$

In other words, that's a formula which is true when x is equal to 2, and false everywhere else, so it singles out 2?

"Exactly. And this is a formula which detects odd numbers:"

$$\exists y: x=(2*y)+1$$

Um... okay. That formula says, 'There exists a y, such that x equals 2 times y plus one.' And that's true when x is 1, because 0 is a number, and 1=(2*0)+1. And it's true when x is 9, because there exists a number 4 such that 9=(2*4)+1... right. The formula is true at all odd numbers, and only odd numbers.

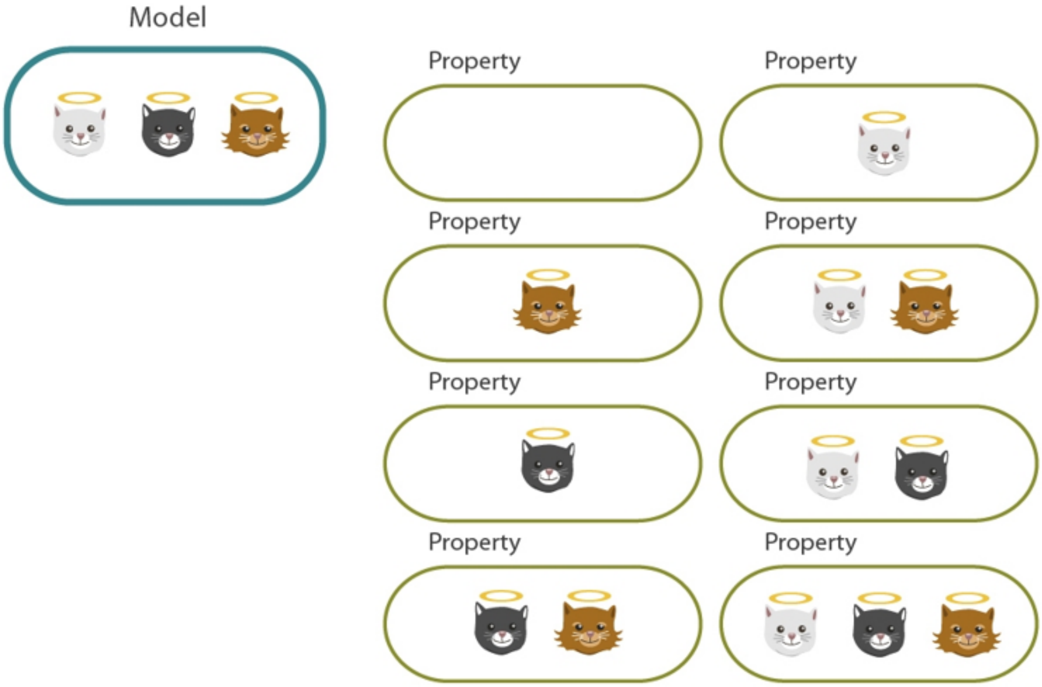
"Indeed. Now suppose we had some way to *detect the existence* of that ABC-loop in the model - a formula which was *true* at the ABC-loop and *false* everywhere else. Then I could adapt the *negation* of this statement to say 'No objects like this are allowed to exist', and add that as an axiom alongside 'Every number has a successor' and so on. Then I'd have *narrowed down* the possible set of models to get rid of models that have an extra ABC-loop in them."

Um... can I rule out the ABC-loop by saying ¬∃x:(x=A)?

"Er, only if you've told me what A is in the first place, and in a logic which has ruled out all models with loops in them, you shouldn't be able to point to a specific object that doesn't exist -"

Right. Okay... so the idea is to rule out loops of successors... hm. In the numbers 0, 1, 2, 3..., the number 0 isn't the successor of any number. If I just took a group of numbers starting at 1, like {1, 2, 3, ...}, then 1 wouldn't be the successor of any number *inside* that group. But in A, B, C, the number A is the successor of C, which is the successor of B, which is the successor of A. So how about if I say: 'There's no group of numbers G such that for any number x in G, x is the successor of some other number y in G.'

"Ah! Very clever. But it so happens that you just used second-order logic, because you talked about *groups* or *collections* of entities, whereas *first-order logic* only talks about *individual* entities. Like, suppose we had a logic talking about kittens and whether they're innocent. Here's a model of a universe containing exactly three distinct kittens who are all innocent:"



Er, what are those 'property' thingies?

"They're all possible collections of kittens. They're labeled *properties* because every collection of kittens corresponds to a property that some kittens have and some kittens don't. For example, the collection on the top right, which contains only the grey kitten, corresponds to a predicate which is true at the grey kitten and false everywhere else, or to a property which the grey kitten has which no other kitten has. Actually, for now let's just pretend that 'property' just says 'collection'."

Okay. I understand the concept of a collection of kittens.

"In first-order logic, we can talk about individual kittens, and how they relate to other individual kittens, and whether or not any kitten bearing a certain relation exists or doesn't exist. For example, we can talk about how the grey kitten adores the brown kitten. In second-order logic, we can talk about collections of kittens, and whether or not those collections exist. So in first-order logic, I can say, 'There exists a kitten which is innocent', or 'For every individual kitten, that kitten is innocent', or 'For every individual kitten, there exists another individual kitten which adores the first kitten.' But it requires second-order logic to make statements about *collections* of kittens, like, 'There exists no collection of kittens such that every kitten in it is adored by some other kitten inside the collection.'"

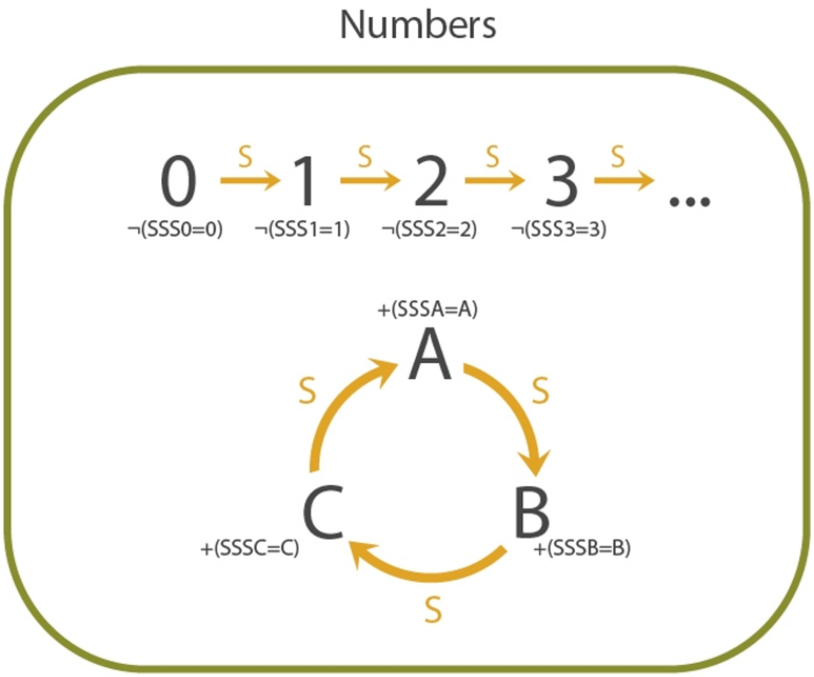
I see. So when I tried to say that you couldn't have any group of numbers, such that every number in the group was a successor of some other number in the group...

"...you quantified over the existence or nonexistence of *collections* of numbers, which means you were using *second-order logic.* However, in this particular case, it's easily possible to rule out the ABC-loop of numbers using only first-order logic. Consider the formula:"

$x=SSSx$

x plus 3 is equal to itself?

"Right. That's a first-order formula, since it doesn't talk about collections. And that formula is false at 0, 1, 2, 3... but true at A, B, and C."



What does the '+' mean?

"Er, by '+' I was trying to say, 'this formula works out to True' and similarly '¬' was supposed to mean the formula works out to False. The general idea is that we now have a formula for detecting 3-loops, and distinguishing them from *standard* numbers like 0, 1, 2 and so on."

I see. So by adding the new axiom, ¬∃x:x=SSSx, we could rule out all the models containing A, B, and C or any other 3-loop of nonstandard numbers.

"Right."

But this seems like a rather arbitrary sort of axiom to add to a fundamental theory of arithmetic. I mean, I've never seen any attempt to describe the numbers which says, 'No number is equal to itself plus 3' as a basic premise. It seems like it should be a theorem, not an axiom.

"That's because it's brought in using a more general rule. In particular, first-order arithmetic has an *infinite axiom schema* - an infinite but computable scheme of axioms. Each axiom in the schema says, for a different first-order formula Φ(x) - pronounced 'phi of x' - that:"

1. *If* Φ is true at 0, i.e: Φ(0)
2. *And if* Φ is true of the successor of any number where it's true, i.e:
   ∀x: Φ(x)→Φ(Sx)
3. *Then* Φ is true of all numbers: ∀n: Φ(n)

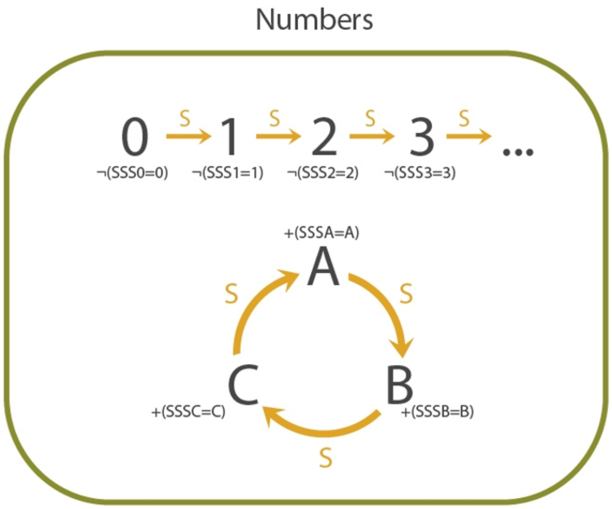
$$(\Phi(0) \wedge (\forall x: \Phi(x) \rightarrow \Phi(Sx))) \rightarrow (\forall n: \Phi(n))$$

"In other words, every *formula* which is true at 0, and which is true of the successor of any number of which it is true, is true *everywhere.* This is the *induction schema* of first-order arithmetic. As a special case we have the *particular* inductive axiom:"

$$(0 \neq SSS0 \wedge (\forall x: (x \neq SSSx) \rightarrow (Sx \neq SSSSx)) \rightarrow (\forall n: n \neq SSSn)$$

But that doesn't say that for all n, $n \neq n+3$. It gives some premises from which that conclusion would follow, but we don't know the premises.

"Ah, however, we can *prove* those premises using the *other* axioms of arithmetic, and hence prove the conclusion. The formula $(SSSx=x)$ is false at 0, because 0 is not the successor of *any* number, including SS0. Similarly, consider the formula $SSSSx=Sx$, which we can rearrange as $S(SSSx)=S(x)$. If two numbers have the same successor they are the same number, so $SSSx=x$. If truth at Sx proves truth at x, then falsity at x proves falsity at Sx, modus ponens to modus tollens. Thus the formula is false at zero, false of the successor of any number where it's false, and so must be false everywhere under the induction axiom schema of first-order arithmetic. And so first-order arithmetic can rule out models like this:"

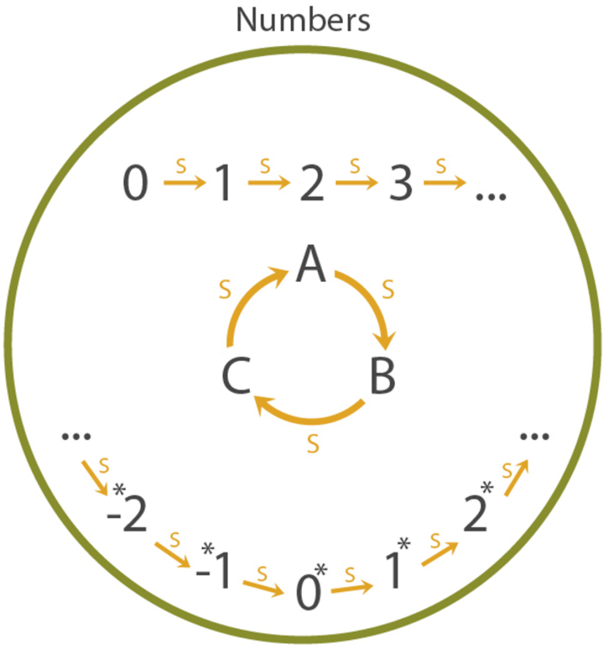


...er, I think I see? Because if this model obeys all the *other* axioms which which we *already* specified, that *didn't* filter it out earlier - axioms like 'zero is not the successor of any number' and 'if two numbers have the same successor they are the same number' - then we can *prove* that the formula $x \neq SSSx$ is true at 0, and prove that if the formula true at x it must be true at x+1. So once we then add the *further* axiom that *if* $x \neq SSSx$ is true at 0, and *if* $x \neq SSSx$ is true at Sy when it's true at y, *then* $x \neq SSSx$ is true at all x...

"We already have the premises, so we get the conclusion. $\forall x: x \neq SSSx$, and thus we filter out all the 3-loops. Similar logic rules out N-loops for all N."

So then did we get rid of all the nonstandard numbers, and leave only the standard model?

"No. Because there was also that problem with the infinite chain ... -2*, -1*, 0*, 1* and so on."



Here's one idea for getting rid of the model with an infinite chain. All the nonstandard numbers in the chain are "greater" than all the standard numbers, right? Like, if *w* is a nonstandard number, then *w* > 3, *w* > 4, and so on?

"Well, we can prove by induction that no number is less than 0, and *w* isn't equal to 0 or 1 or 2 or 3, so I'd have to agree with that."

Okay. We should also be able to prove that if x > y then x + z > y + z. So if we take nonstandard *w* and ask about *w* + *w*, then *w* + *w* must be greater than *w* + 3, *w* + 4, and so on. So *w* + *w* can't be part of the infinite chain at all, and yet adding any two numbers ought to yield a third number.

"Indeed, that does prove that if there's one infinite chain, there must be *two* infinite chains. In other words, that original, exact model in the picture, can't all by itself be a model of first-order arithmetic. But showing that the chain implies the existence of yet other elements, isn't the same as proving that the chain doesn't exist. Similarly, since all numbers are even or odd, we must be able to find *v* with *v* + *v* = *w*, or find *v* with *v* + *v* + 1 = *w*. Then *v* must be part of another nonstandard chain that comes before the chain containing *w*."

But then that requires an *infinite* number of infinite chains of nonstandard numbers which are all greater than any standard number. Maybe we can extend this logic to eventually reach a contradiction and rule out the existence of an infinite chain in the first place - like, we'd show that any complete collection of nonstandard numbers has to be *larger than itself* -

"Good idea, but no. You end up with the conclusion that if a single nonstandard number exists, it must be part of a chain that's infinite in both directions, i.e., a chain that looks like an ordered copy of the negative and positive integers. And that if an infinite chain exists, there must be infinite chains corresponding to all *rational numbers.* So something that could actually be a nonstandard model of first-order arithmetic, has to contain at least the standard numbers *followed by* a copy of the rational numbers with each rational number replaced by a copy of the integers. But then *that* setup works just fine with both addition and multiplication - we can't prove that it has to be any larger than what we've already said."

Okay, so how *do* we get rid of an infinite number of infinite chains of nonstandard numbers, and leave just the standard numbers at the beginning? What kind of statement would they violate - what sort of axiom would rule out all those extra numbers?

"We have to use second-order logic for that one."

Honestly I'm still not 100% clear on the difference.

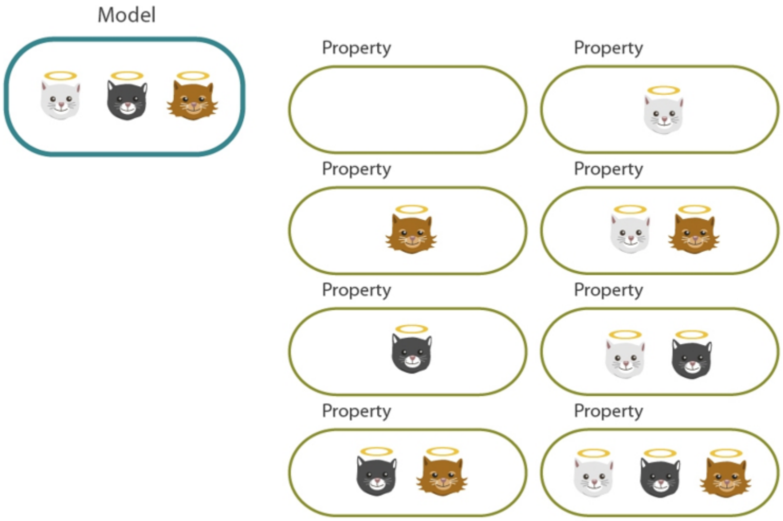
"Okay... earlier you gave me a *formula* which detected odd numbers."

Right. ∃y: x=(2*y)+1, which was true at x=1, x=9 and so on, but not at x=0, x=4 and so on.

"When you think in terms of *collections of numbers,* well, there's *some* collections which can be defined by formulas. For example, the collection of odd numbers {1, 3, 5, 7, 9, ...} can be defined by the formula, with x free, ∃y: x=(2*y)+1. But you could also try to talk about just the collection {1, 3, 5, 7, 9, ...} as a collection, a set of numbers, whether or not there happened to be any formula that defined it -"

Hold on, how can you talk about a set if you can't define a formula that makes something a member or a non-member? I mean, that seems a bit smelly from a rationalist perspective -

"Er... remember the earlier conversation about kittens?"



"Suppose you say something like, 'There *exists a collection* of kittens, such that every kitten adores only other kittens in the collection'. Give me a room full of kittens, and I can count through all possible collections, check your statement for each collection, and see whether or not there's a collection which is actually like that. So the statement is meaningful - it can be falsified or verified, and it constrains the state of reality. But you didn't give me a *local formula* for picking up a *single* kitten and deciding whether or not it ought to be in this mysterious collection. I had to iterate through all the *collections* of kittens, find the *collections* that matched your statement, and only then could I decide whether any individual kitten had the property of being in a collection like that. But the statement was still falsifiable, even though it was, in mathematical parlance, *impredicative* - that's what we call it when you make a statement that can only be verified by looking at many possible collections, and doesn't start from any particular collection that you tell me how to construct."

Ah... hm. What about infinite universes of kittens, so you can't iterate through all possible collections in finite time?

"If you say, 'There exists a collection of kittens which all adore each other', I could exhibit a group of three kittens which adored each other, and so prove the statement true. If you say 'There's a collection of four kittens who adore only each other', I might come up with a constructive proof, given the other known properties of kittens, that your statement was false; and any time you tried giving me a group of four kittens, I could find a fifth kitten, adored by some kitten in your group, that falsified your attempt. But this is getting us into some rather deep parts of math we should probably stay out of for now. The point is that even in infinite universes, there are second-order statements that you can prove or falsify in finite amounts of time. And once you admit those *particular* second-order statements are talking about something meaningful, well, you might as well just admit that second-order statements in general are meaningful."

...that sounds a little iffy to me, like we might get in trouble later on.

"You're not the only mathematician who worries about that."

But let's get back to numbers. You say that we can use second-order logic to rule out any infinite chain.

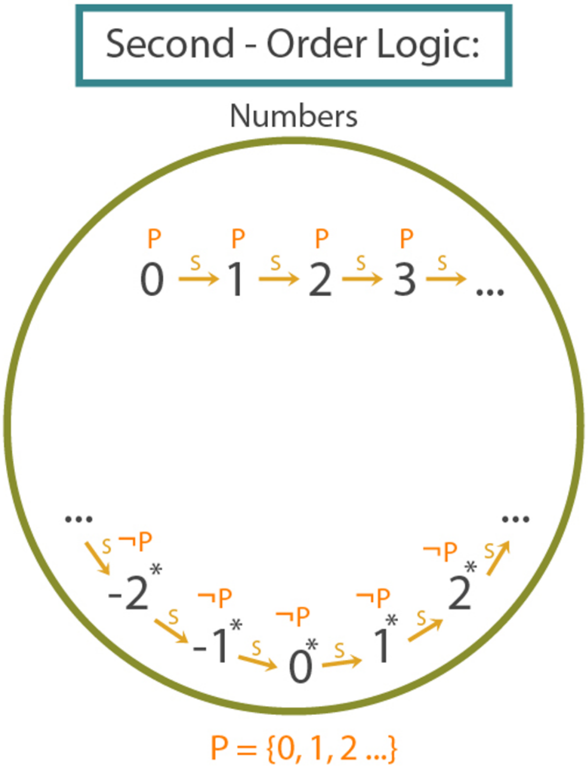
"Indeed. In second-order logic, instead of using an infinite axiom schema over all formulas Φ, we quantify over *possible collections* directly, and say, in a *single* statement:"

$$\forall P: P(0) \wedge (\forall x: P(x) \rightarrow P(Sx)) \rightarrow (\forall n: P(n))$$

"Here P is any predicate true or false of individual numbers. Any collection of numbers corresponds to a predicate that is true of numbers inside the collection and false of numbers outside of it."

Okay... and how did that rule out infinite chains again?

"Because *in principle,* whether or not there's any first-order formula that picks them out, there's *theoretically* a collection that contains the standard numbers {0, 1, 2, ...} and *only* the standard numbers. And if you treat that collection as a predicate P, then P is true at 0 - that is, 0 is in the standard numbers. And if 200 is a standard number then so is 201, and so on; if P is true at x, it's true at x+1. On the other hand, if you treat the collection 'just the standard numbers' as a predicate, it's false at -2*, false at -1*, false at 0* and so on - those numbers *aren't* in this theoretical collection. So it's vacuously true that this predicate is true at 1* if it's true at 0*, because it's *not* true at 0*. And so we end up with:"



"And so the single second-order axiom..."

$$\forall P: P0 \land (\forall x: Px \rightarrow P(Sx)) \rightarrow (\forall n: Pn)$$

"...rules out any disconnected chains, finite loops, and indeed every nonstandard number, in one swell foop."

But what did that axiom *mean,* exactly? I mean, taboo the phrase 'standard numbers' for a moment, pretend I've got no idea what those are, just explain to me what the axiom actually *says*.

"It says that the model being discussed - the model which fits this axiom - makes it impossible to form *any collection closed under succession* which includes 0 and doesn't include *everything*. It's impossible to have *any collection of objects in this universe* such that 0 is in the collection, and the successor of everything in the collection is in the collection, and yet this collection doesn't contain *everything.* So you can't have a disconnected infinite chain - there would then exist at least one collection over objects in this universe that contained 0 and all its successor-descendants, yet didn't contain the chain; and we have a shiny new axiom which says that can't happen."

Can you perhaps operationalize that in a more sensorymotory sort of way? Like, if this is what I believe about the universe, then what do I expect to see?

"If this is what you believe about the mathematical model that you live in... then you believe that neither you, nor any adversary, nor yet a superintelligence, nor yet God, can consistently say 'Yea' or 'Nay' to objects in such fashion that when you present them with 0, they say 'Yea', and when you present them with any other object, if they say 'Yea', they also say 'Yea' for the successor of that object; and yet there is some object for which they say 'Nay'. You believe this can never happen, no matter what. The way in which the objects in the universe are arranged by succession, just doesn't let that happen, ever."

Ah. So if, say, they said 'Nay' for 42, I'd go back and ask about 41, and then 40, and by the time I reached 0, I'd find either that they said 'Nay' about 0, or that they said 'Nay' for 41 and yet 'Yea' for 40. And what do I expect to see if I believe in first-order arithmetic, with the infinite axiom schema?

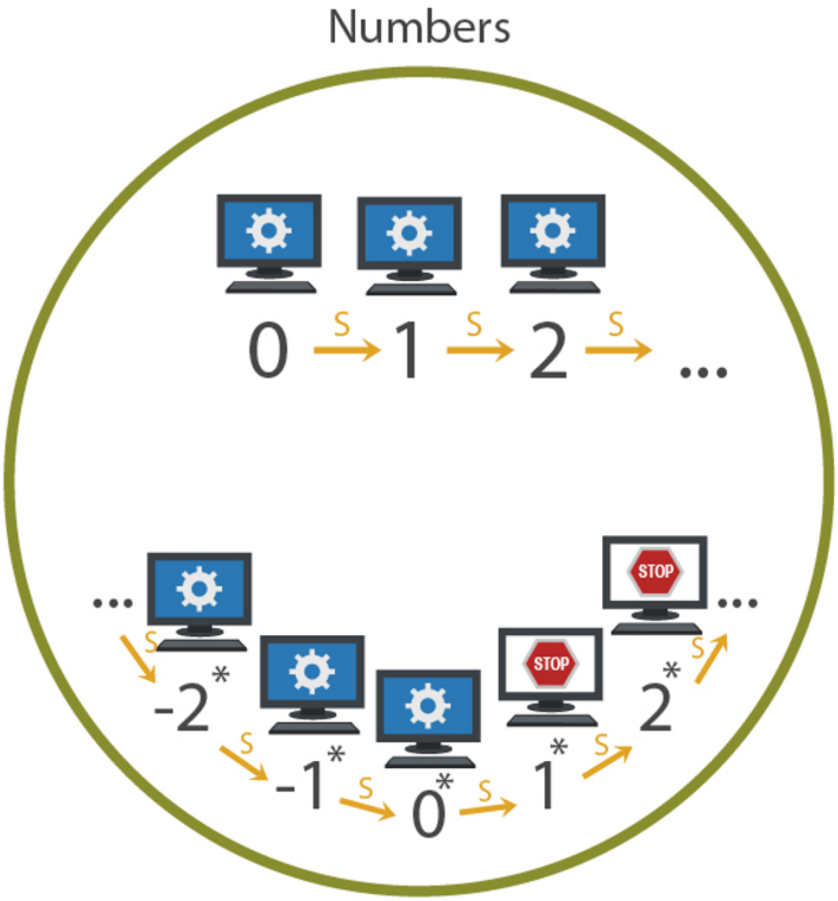
"In that case, you believe there's no neatly specifiable, compactly describable *rule* which behaves like that. But if you believe the second-order version, you believe nobody can possibly behave like that even if they're answering randomly, or branching the universe to answer different ways in different alternate universes, and so on. And note, by the way, that if we have a finite universe - i.e., we throw out the rule that *every* number has a successor, and say instead that 256 is the only number which has no successor - then we can verify this axiom in finite time."

I see. Still, is there any way to rule out infinite chains using *first*-order logic? I might find that easier to deal with, even if it looks more complicated at first.

"I'm afraid not. One way I like to look at it is that first-order logic can talk about *constraints on how the model looks from any local point*, while only second-order logic can talk about *global qualities* of chains, collections, and the model as a whole. Whether every number has a successor is a local property - a question of how the model looks from the vantage point of any one number. Whether a number plus three, can be equal to itself, is a question you could evaluate at the local vantage point of any one number. Whether a number is *even,* is a question you can answer by looking around for a single, individual number x with the property that x+x equals the first number. But when you try to say that there's *only one connected chain* starting at 0, by invoking the idea of *connectedness* and *chains* you're trying to describe non-local properties that require a logic-of-possible-collections to specify."

Huh. But if all the 'local' properties are the same regardless, why worry about global properties? In first-order arithmetic, any 'local' formula that's true at zero and all of its 'natural' successors would also have to be true of all the disconnected infinite chains... right? Or did I make an error there? All the other infinite chains besides the 0-chain - all 'nonstandard numbers' - would have just the same properties as the 'natural' numbers, right?

"I'm afraid not. The first-order axioms of arithmetic may fail to pin down whether or not a Turing machine halts - whether there *exists a time* at which a Turing machine halts. Let's say that from our perspective inside the standard numbers, the Turing machine 'really doesn't' halt - it doesn't halt on clock tick 0, doesn't halt on clock tick 1, doesn't halt on tick 2, and so on through all the standard successors of the 0-chain. In nonstandard models of the integers - models with other infinite chains - there might be somewhere inside a *nonstandard chain* where the Turing machine goes from running to halted and stays halted thereafter."



"In this new model - which is fully compatible with the first-order axioms, and can't be ruled out by them - it's not true that 'for every number t at which the Turing machine is running, it will still be running at t+1'. Even though if we could somehow restrict our attention to the 'natural' numbers, we would see that the Turing machine was running at 0, 1, 2, and every time in the successor-chain of 0."

Okay... I'm not quite sure what the *practical* implication of that is?

"It means that many Turing machines which *in fact* never halt at any standard time, can't be *proven not to halt* using first-order reasoning, because their non-halting-ness *does not actually follow logically* from the first-order axioms. Logic is about which conclusions follow from which premises, remember? If there are models which are compatible with all the first-order premises, but still falsify the statement 'X runs forever', then the statement 'X runs forever' can't *logically follow* from those premises. This means you won't be able to prove - *shouldn't* be able to prove - that this Turing machine halts, using *only* first-order logic."

How exactly would this fail in practice? I mean, where does the proof go bad?

"You wouldn't get the second step of the induction, 'for every number t at which the Turing machine is running, it will still be running at t+1'. There'd be nonstandard models with some nonstandard t that falsifies the premise - a nonstandard time where the Turing machine goes from running to halted. Even though if we could somehow restrict our attention to *only the standard numbers*, we would see that the Turing machine was running at 0, 1, 2, and so on."

But if a Turing machine really actually halts, there's got to be some *particular time* when it halts, like on step 97 -

"Indeed. But 97 exists in *all* nonstandard models of arithmetic, so we can prove its existence in first-order logic. Any time 0 is a number, every number has a successor, numbers don't loop, and so on, there'll exist 97. Every nonstandard model has *at least* the standard numbers. So whenever a Turing machine *does* halt, you can prove in first-order arithmetic that it halts - it does indeed follow from the premises. That's kinda what you'd *expect,* given that you can just watch the Turing machine for 97 steps. When something actually does halt, you *should* be able to prove it halts without worrying about unbounded future times! It's when something *doesn't actually* halt - in the standard numbers, that is - that the existence of 'nonstandard halting times' becomes a problem. Then, the conclusion that the Turing machine runs forever *may not actually follow* from first-order arithmetic, because you can obey all the premises of first-order arithmetic, and yet still be inside a nonstandard model where this Turing machine halts at a nonstandard time."

So second-order arithmetic is more powerful than first-order arithmetic in terms of *what follows from the premises*?

"That follows inevitably from the ability to talk about *fewer possible models*. As it is written, 'What is true of one apple may not be true of another apple; thus more can be said about a single apple than about all the apples in the world.' If you can restrict your discourse to a narrower collection of models, there are more facts that follow inevitably, because the more models you might be talking about, the fewer facts can possibly be true about all of them. And it's also definitely true that second-order arithmetic proves more theorems than first-order arithmetic - for example, it can prove that a Turing machine which computes Goodstein sequences always reaches 0 and halts, or that Hercules always wins the hydra game. But there's a bit of controversy we'll get into later about whether second-order logic is *actually* more powerful than first-order logic in general."

Well, sure. After all, just because nobody has ever yet invented a first-order formula to filter out all the nonstandard numbers, doesn't mean it can never, ever be done. Tomorrow some brilliant mathematician might figure out a way to take an individual number x, and do local things to it using addition and multiplication and the existence or nonexistence of other individual numbers, which can tell us whether that number is part of the 0-chain or some other infinite-in-both-directions chain. It'll be as easy as (a=b*c) -

"Nope. Ain't never gonna happen."

But maybe you could find some entirely different creative way of first-order axiomatizing the numbers which has *only* the standard model -

"Nope."

Er... how do you *know* that, exactly? I mean, part of the Player Character Code is that you don't give up when something *seems* impossible. I can't quite see *yet* how to detect infinite chains using a first-order formula. But then earlier I didn't realize you could rule out finite loops, which turned out to be quite simple once you explained. After all, there's two distinct uses of the word 'impossible', one which indicates positive knowledge that something can *never* be done, that no *possible* chain of actions can *ever* reach a goal, even if you're a superintelligence. This kind of knowledge requires a strong, definite grasp on the subject matter, so that you can rule out *every* possible avenue of success. And then there's another, *much more common* use of the word 'impossible', which means that you thought about it for five seconds but didn't see any way to do it, usually used in the presence of *weak* grasps on a subject, subjects that seem sacredly mysterious -

"Right. Ruling out an infinite-in-both-directions chain, using a first-order formula, is the *first* kind of impossibility. We *know* that it can never be done."

I see. Well then, what do you think you know, and how do you think you know it? How is this definite, positive knowledge of impossibility obtained, using your strong grasp on the non-mysterious subject matter?

"We'll take that up next time."

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

Next post: "Godel's Completeness and Incompleteness Theorems"

Previous post: "By Which It May Be Judged"

Ontological Crisis in Humans

Imagine a robot that was designed to find and collect spare change around its owner's house. It had a world model where macroscopic everyday objects are ontologically primitive and ruled by high-school-like physics and (for humans and their pets) rudimentary psychology and animal behavior. Its goals were expressed as a utility function over this world model, which was sufficient for its designed purpose. All went well until one day, a prankster decided to "upgrade" the robot's world model to be based on modern particle physics. This unfortunately caused the robot's utility function to instantly throw a domain error exception (since its inputs are no longer the expected list of macroscopic objects and associated properties like shape and color), thus crashing the controlling AI.

According to Peter de Blanc, who used the phrase "ontological crisis" to describe this kind of problem,

> Human beings also confront ontological crises. We should find out what cognitive algorithms humans use to solve the same problems described in this paper. If we wish to build agents that maximize human values, this may be aided by knowing how humans re-interpret their values in new ontologies.

I recently realized that a couple of problems that I've been thinking over (the nature of selfishness and the nature of pain/pleasure/suffering/happiness) can be considered instances of ontological crises in humans (although I'm not so sure we necessarily have the cognitive algorithms to solve them). I started thinking in this direction after writing this comment:

> This formulation or variant of TDT requires that before a decision problem is handed to it, the world is divided into the agent itself (X), other agents (Y), and "dumb matter" (G). I think this is misguided, since the world doesn't really divide cleanly into these 3 parts.

What struck me is that even though the world doesn't divide cleanly into these 3 parts, *our models* of the world actually do. In the world models that we humans use on a day to day basis, and over which our utility functions seem to be defined (to the extent that we can be said to have utility functions at all), we do take the Self, Other People, and various Dumb Matter to be ontologically primitive entities. Our world models, like the coin collecting robot's, consist of these macroscopic objects ruled by a hodgepodge of heuristics and prediction algorithms, rather than microscopic particles governed by a coherent set of laws of physics.

For example, the amount of pain someone is experiencing doesn't seem to exist in the real world as an XML tag attached to some "person entity", but that's pretty much how our models of the world work, and perhaps more importantly, that's what our utility functions expect their inputs to look like (as opposed to, say, a list of particles and their positions and velocities). Similarly, a human can be selfish just by treating the object labeled "SELF" in its world model differently from other objects, whereas an AI with a world model consisting of microscopic particles would need to somehow inherit or learn a detailed description of itself in order to be selfish.

To fully confront the ontological crisis that we face, we would have to upgrade our world model to be based on actual physics, and simultaneously translate our utility functions so that their domain is the set of possible states of the new model. We currently have little idea how to accomplish this, and instead what we do in practice is, as far as I can tell, keep our ontologies intact and utility functions unchanged, but just add some new heuristics that in certain limited circumstances call out to new physics formulas to better update/extrapolate our models. This is actually rather clever, because it lets us make use of updated understandings of physics without ever having to, for instance, decide exactly what patterns of particle movements constitute pain or pleasure, or what patterns constitute oneself. Nevertheless, this approach hardly seems capable of being extended to work in a future where many people may have nontraditional mind architectures, or have a zillion copies of themselves running on all kinds of strange substrates, or be merged into amorphous group minds with no clear boundaries between individuals.

By the way, I think nihilism often gets short changed around here. Given that we do not actually have at hand a solution to ontological crises in general or to the specific crisis that we face, what's wrong with saying that the solution set may just be null? Given that evolution doesn't constitute a particularly benevolent and farsighted designer, perhaps we may not be able to do much better than that poor spare-change collecting robot? If Eliezer is worried that actual AIs facing actual ontological crises could do worse than just crash, should we be very sanguine that for humans everything must "add up to moral normality"?

To expand a bit more on this possibility, many people have an aversion against moral arbitrariness, so we need at a minimum a utility translation scheme that's principled enough to pass that filter. But our existing world models are a hodgepodge put together by evolution so there may not be any such sufficiently principled scheme, which (if other approaches to solving moral philosophy also don't pan out) would leave us with legitimate feelings of "existential angst" and nihilism. One could perhaps still argue that any *current* such feelings are premature, but maybe some people have stronger intuitions than others that these problems are unsolvable?

Do we have any examples of humans successfully navigating an ontological crisis? The LessWrong Wiki mentions loss of faith in God:

> In the human context, a clear example of an ontological crisis is a believer’s loss of faith in God. Their motivations and goals, coming from a very specific view of life suddenly become obsolete and maybe even nonsense in the face of this new configuration. The person will then experience a deep crisis and go through the psychological task of reconstructing its set of preferences according the new world view.

But I don't think loss of faith in God actually constitutes an ontological crisis, or if it does, certainly not a very severe one. An ontology consisting of Gods, Self, Other People, and Dumb Matter just isn't very different from one consisting of Self, Other People, and Dumb Matter (the latter could just be considered a special case of the former with quantity of Gods being 0), especially when you compare either ontology to one made of microscopic particles or even less familiar entities.

But to end on a more positive note, realizing that seemingly unrelated problems are actually instances of a more general problem gives some hope that by "going meta" we can find a solution to all of these problems at once. Maybe we can solve many ethical problems simultaneously by discovering some generic algorithm that can be used by an agent to transition from any ontology to another?

(Note that I'm not saying this *is* the right way to understand one's real preferences/morality, but just drawing attention to it as a possible alternative to other more "object level" or "purely philosophical" approaches. See also this previous discussion, which I recalled after writing most of the above.)

Godel's Completeness and Incompleteness Theorems

Followup to: Standard and Nonstandard Numbers

So... last time you claimed that using first-order axioms to rule out the existence of nonstandard numbers - other chains of numbers besides the 'standard' numbers starting at 0 - was *forever and truly impossible*, even unto a superintelligence, no matter *how* clever the first-order logic used, even if you came up with an entirely different way of axiomatizing the numbers.

"Right."

How could you, in your finiteness, possibly know that?

"Have you heard of Godel's Incompleteness Theorem?"

Of course! Godel's Theorem says that for every consistent mathematical system, there are statements which are *true* within that system, which can't be *proven* within the system itself. Godel came up with a way to encode theorems and proofs as numbers, and wrote a purely numerical formula to detect whether a proof obeyed proper logical syntax. The basic trick was to use prime factorization to encode lists; for example, the ordered list <3, 7, 1, 4> could be uniquely encoded as:

$2^3 * 3^7 * 5^1 * 7^4$

And since prime factorizations are unique, and prime powers don't mix, you could inspect this single number, 210,039,480, and get the unique ordered list <3, 7, 1, 4> back out. From there, going to an encoding for logical formulas was easy; for example, you could use the 2 prefix for NOT and the 3 prefix for AND and get, for any formulas Φ and Ψ encoded by the numbers $\#\Phi$ and $\#\Psi$:

$\neg\Phi = 2^2 * 3^{\#\Phi}$

$\Phi \wedge \Psi = 2^3 * 3^{\#\Phi} * 5^{\#\Psi}$

It was then possible, by dint of crazy amounts of work, for Godel to come up with a gigantic formula of Peano Arithmetic [](p, c) meaning, 'P encodes a valid logical proof using first-order Peano axioms of C', from which directly followed the formula []c, meaning, 'There exists a number P such that P encodes a proof of C' or just 'C is provable in Peano arithmetic.'

Godel then put in some *further* clever work to invent statements which referred to *themselves*, by having them contain sub-recipes that would reproduce the entire statement when manipulated by another formula.

And then Godel's Statement encodes the statement, 'There does not exist any number P such that P encodes a proof of (this statement) in Peano arithmetic' or in simpler terms 'I am not provable in Peano arithmetic'. If we assume first-order arithmetic is consistent and sound, then no *proof* of this statement *within* first-order arithmetic exists, which means the statement is *true* but can't be proven within the system. That's Godel's Theorem.

"Er... no."

No?

"No. I've heard rumors that Godel's Incompleteness Theorem is horribly misunderstood in your Everett branch. Have you heard of Godel's *Completeness* Theorem?"

Is that a thing?

"Yes! Godel's Completeness Theorem says that, for any collection of first-order statements, *every semantic implication of those statements is syntactically provable within first-order logic*. If something is a genuine implication of a collection of first-order statements - if it actually *does* follow, in the models pinned down by those statements - then you can *prove* it, *within* first-order logic, using *only* the syntactical rules of proof, from those axioms."

I don't see how that could possibly be true at the same time as Godel's Incompleteness Theorem. The Completeness Theorem and Incompleteness Theorem seem to say diametrically opposite things. Godel's Statement is implied by the axioms of first-order arithmetic - that is, we can see it's true using our own mathematical reasoning -

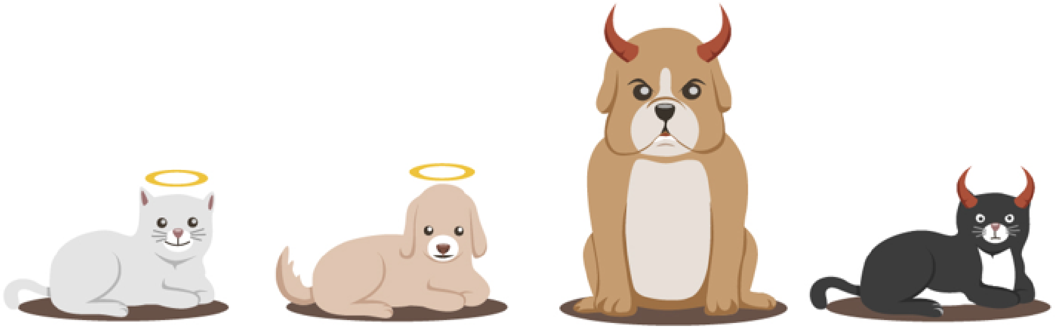
"Wrong."

What? I mean, I understand we can't prove it *within* Peano arithmetic, but from outside the system we can see that -

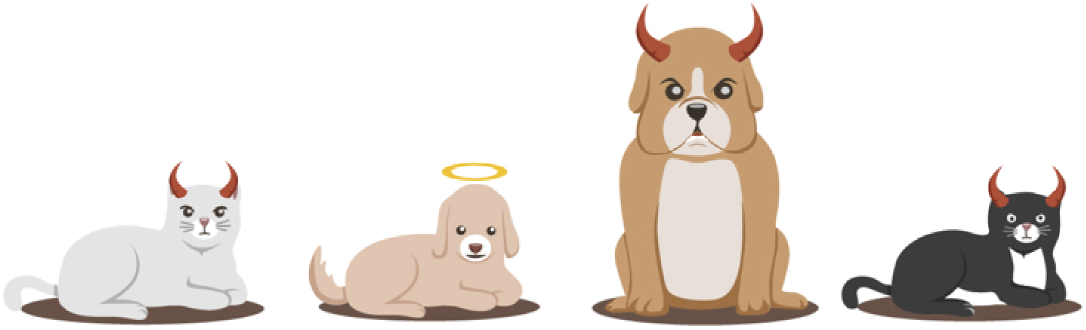
All right, explain.

"Basically, you just committed the equivalent of saying, 'If all kittens are little, and some little things are innocent, then some kittens are innocent.' There are universes -

logical models - where it so happens that the premises are true and the conclusion also happens to be true:"



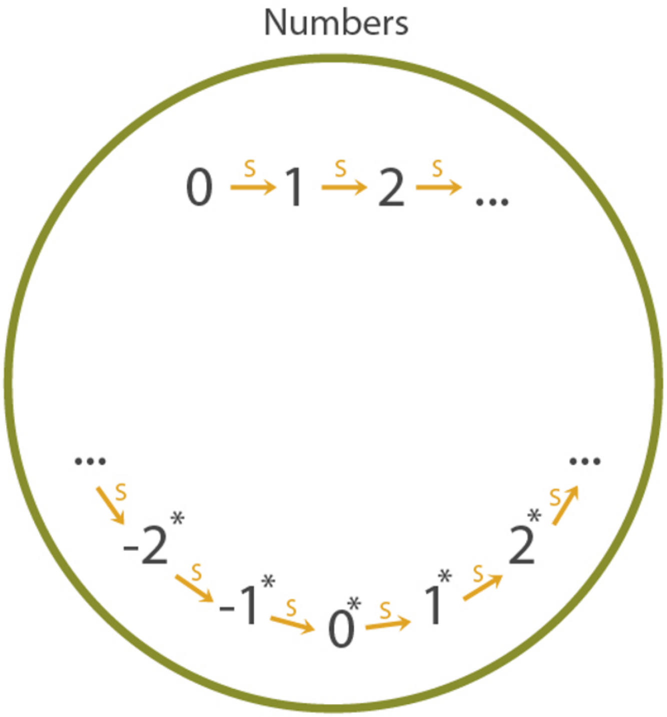
"But there are also valid models of the premises where the conclusion is false:"



"If you, yourself, happened to live in a universe like the first one - if, in your mind, you were *only thinking* about a universe like that - then you might *mistakenly* think that you'd proven the conclusion. But your statement is not *logically* valid, the conclusion is not true in *every* universe where the premises are true. It's like saying, 'All apples are plants. All fruits are plants. Therefore all apples are fruits.' Both the premises and the conclusions happen to be true in *this* universe, but it's not valid logic."

Okay, so how does this invalidate my previous explanation of Godel's Theorem?

"Because of the non-standard models of first-order arithmetic. First-order arithmetic narrows things down a lot - it rules out 3-loops of nonstandard numbers, for example, and mandates that every model contain the number 17 - but it doesn't pin down a *single* model. There's still the possibility of infinite-in-both-directions chains coming after the 'standard' chain that starts with 0. Maybe *you* have just the standard numbers in mind, but that's not the *only* possible model of first-order arithmetic."



So?

"So in some of those other models, there are nonstandard numbers which - according to Godel's *arithmetical* formula for encodes-a-proof - are 'nonstandard proofs' of Godel's Statement. I mean, they're not what we would call *actual* proofs. An actual proof would have a standard number corresponding to it. A nonstandard proof might look like... well, it's hard to envision, but it might be something like, 'Godel's statement is true, because not-not-Godel's statement, because not-not-not-not-Godel's statement', and so on going *backward forever*, every step of the proof being valid, because nonstandard numbers have an infinite number of predecessors."

And there's no way to say, 'You can't have an infinite number of derivations in a proof'?

"Not in first-order logic. If you could say that, you could rule out numbers with infinite numbers of predecessors, meaning that you could rule out all infinite-in-both-directions chains, and hence rule out all nonstandard numbers. And then the only *remaining* model would be the standard numbers. And then Godel's Statement would be a *semantic* implication of those axioms; there would exist *no* number encoding a proof of Godel's Statement in *any* model which obeyed the axioms of first-order arithmetic. And then, by Godel's *Completeness* Theorem, we could prove Godel's Statement from those axioms using first-order syntax. Because every *genuinely* valid implication of any collection of first-order axioms - every first-order statement that *actually does follow, in every possible model where the premises are true* - can *always* be proven, from those axioms, in first-order logic. Thus, by the *combination* of Godel's Incompleteness Theorem and Godel's Completeness Theorem, we see that there's no way to uniquely pin down the natural numbers using first-order logic. QED."

Whoa. So everyone in the human-superiority crowd gloating about how *they're* superior to mere machines and formal systems, because *they* can see that Godel's Statement is true just by their sacred and mysterious mathematical intuition...

"...Is actually committing a horrendous logical fallacy of the sort that no cleanly designed AI could ever be tricked into, yes. Godel's Statement doesn't *actually follow* from the first-order axiomatization of Peano arithmetic! There are models where all the first-order axioms are true, and yet Godel's Statement is false! The standard misunderstanding of Godel's Statement *is* something like the situation as it obtains in *second*-order logic, where there's no equivalent of Godel's Completeness Theorem. But people in the human-superiority crowd usually don't attach that disclaimer - they usually present arithmetic using the first-order version, when they're explaining what it is that they can see that a formal system can't. It's safe to say that *most* of them are inadvertently illustrating the irrational overconfidence of humans jumping to conclusions, even though there's a less stupid version of the same argument which invokes second-order logic."

Nice. But still... that proof you've shown me seems like a rather *circuitous* way of showing that you can't ever rule out infinite chains, especially since I don't see why Godel's Completeness Theorem should be true.

"Well... an equivalent way of stating Godel's Completeness Theorem is that every *syntactically* consistent set of first-order axioms - that is, every set of first-order axioms such that you cannot *syntactically* prove a contradiction from them using first-order logic - has at least one semantic model. The proof proceeds by trying to adjoin statements saying P or ~P for every first-order formula P, at least one of which must be possible to adjoin while leaving the expanded theory syntactically consistent -"

Hold on. Is there some more *constructive* way of seeing why a non-standard model has to exist?

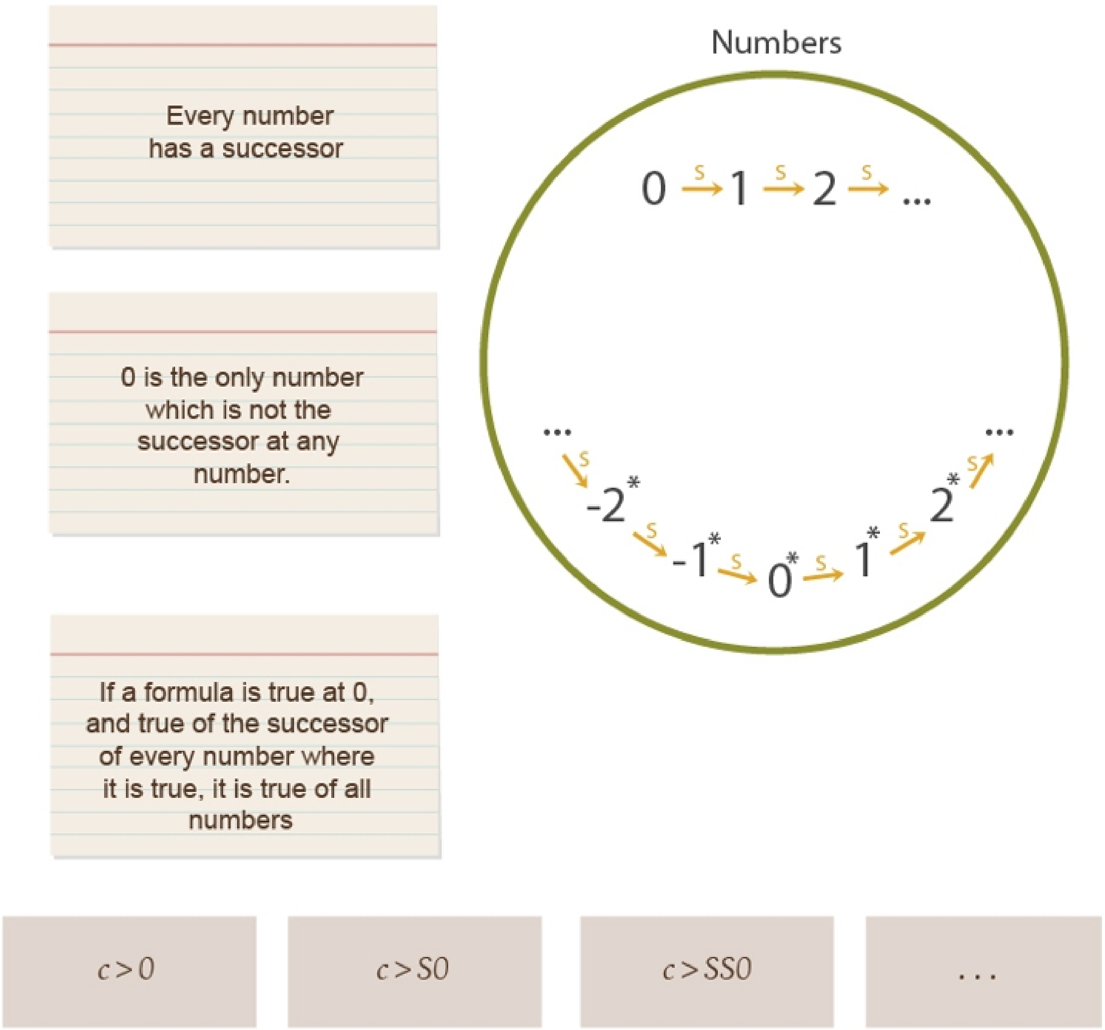
"Mm... you could invoke the Compactness Theorem for first-order logic. The Compactness Theorem says that *if a collection of first-order statements has no model, some finite subset of those statements is also semantically unrealizable*. In other words, if a collection of first-order statements - even an *infinite* collection - is unrealizable in the sense that no possible mathematical model fits all of those premises, then there must be *some* finite subset of premises which are also unrealizable. Or modus ponens to modus tollens, if all finite subsets of a collection of axioms have at least one model, then the whole infinite collection of axioms has at least one model."

Ah, and can you explain why the Compactness Theorem should be true?

"No."

I see.

"But at least it's simpler than the Completeness Theorem, and from the Compactness Theorem, the inability of first-order arithmetic to pin down a standard model of numbers follows immediately. Suppose we take first-order arithmetic, and adjoin an axiom which says, 'There exists a number greater than 0.' Since there does in fact exist a number, 1, which is greater than 0, first-order arithmetic plus this new axiom should be semantically okay - it should have a model if any model of first-order arithmetic ever existed in the first place. Now let's adjoin a new constant symbol *c* to the language, i.e., *c* is a constant symbol referring to a single object across all statements where it appears, the way 0 is a constant symbol and an axiom then identifies 0 as the object which is not the successor of any object. Then we start adjoining axioms saying '*c* is greater than X', where X is some concretely specified number like 0, 1, 17, 2^{256}, and so on. In fact, suppose we adjoin an *infinite* series of such statements, one for every number:"



Every number has a successor

0 is the only number which is not the successor at any number.

If a formula is true at 0, and true of the successor of every number where it is true, it is true of all numbers

Numbers
$0 \xrightarrow{S} 1 \xrightarrow{S} 2 \xrightarrow{S} \ldots$
$\ldots \xrightarrow{S} -2^* \xrightarrow{S} -1^* \xrightarrow{S} 0^* \xrightarrow{S} 1^* \xrightarrow{S} 2^* \xrightarrow{S} \ldots$

$c > 0$

$c > S0$

$c > SS0$

$\ldots$

Wait, so this new theory is saying that there exists a number *c* which is larger than every number?

"No, the infinite schema says that there exists a number *c* which is larger than any *standard* number."

I see, so this new theory *forces* a nonstandard model of arithmetic.

"Right. It rules out *only* the standard model. And the Compactness Theorem says this new theory is still semantically realizable - it has *some* model, just not the standard one."

Why?

"Because any finite subcollection of the new theory's axioms, can only use a finite number of the extra axioms. Suppose the largest extra axiom you used was '*c* is larger than 2^{256}'. In the standard model, there certainly exists a number 2^{256}+1 with which *c* could be consistently identified. So the standard numbers must be a model of that collection of axioms, and thus that finite subset of axioms must be semantically realizable. Thus by the Compactness Theorem, the full, infinite axiom system must also be semantically realizable; it must have at least one model. Now, adding axioms never *increases* the number of compatible models of an axiom system - each additional axiom can only *filter out* models, not *add* models which are incompatible with the other axioms. So this new model of the larger axiom system - containing a number which is greater than 0, greater than 1, and greater than every other 'standard' number - must *also* be a model of first-order Peano arithmetic. That's a relatively simpler proof that first-order arithmetic - in fact, *any* first-order axiomatization of arithmetic - has nonstandard models."

Huh... I can't quite say that seems obvious, because the Compactness Theorem doesn't feel obvious; but at least it seems more specific than trying to prove it using Godel's Theorem.

"A similar construction to the one we used above - adding an infinite series of axioms saying that a thingy is even larger - shows that if a first-order theory has models of unboundedly large finite size, then it has at least one infinite model. To put it even more alarmingly, there's no way to characterize the property of *finiteness* in first-order logic! You can have a first-order theory which characterizes models of cardinality 3 - just say that there exist x, y, and z which are not equal to each other, but with all objects being equal to x or y or z. But there's no first-order theory which characterizes the property of *finiteness* in the sense that all finite models fit the theory, and no infinite model fits the theory. A first-order theory either limits the size of models to some particular upper bound, or it has infinitely large models."

So you can't even say, 'x is finite', without using second-order logic? Just forming the *concept* of infinity and distinguishing it from finiteness requires second-order logic?

"Correct, for pretty much exactly the same reason you can't say 'x is only a finite number of successors away from 0'. You can say, 'x is less than a googolplex' in first-order logic, but not, in full generality, 'x is finite'. In fact there's an even *worse* theorem, the Lowenheim-Skolem theorem, which roughly says that if a first-order theory has *any* infinite model, it has models *of all possible infinite cardinalities.* There are uncountable models of first-order Peano arithmetic. There are countable models of first-order real arithmetic - countable models of any attempt to axiomatize the real numbers in first-order logic. There are countable models of Zermelo-Frankel set theory."

How could you *possibly* have a countable model of the real numbers? Didn't Cantor *prove* that the real numbers were uncountable? Wait, let me guess, Cantor implicitly used second-order logic somehow.

"It follows from the Lowenheim-Skolem theorem that he must've. Let's take Cantor's proof as showing that you can't map every set of integers onto a distinct integer - that is, the powerset of integers is larger than the set of integers. The Diagonal Argument is that if you show me a mapping like that, I can take the set which contains 0 if and only if 0 is not in the set mapped to the integer 0, contains 1 if and only if 1 is *not* in the set mapped to the integer 1, and so on. That gives you a set of integers that no integer maps to."

You know, when I was very young indeed, I thought I'd found a *counterexample* to Cantor's argument. Just take the base-2 integers - 1='1', 2='10', 3='11', 4='100', 5='101', and so on, and let each integer correspond to a set in the obvious way, keeping in mind that I was also young enough to think the integers started at 1:

1	10	11	100	101	110	111	1000	1001
{1}	{2}	{2, 1}	{3}	{3, 1}	{3, 2}	{3, 2, 1}	{4}	{4, 1}

Clearly, every set of integers would map onto a unique integer this way.

"Heh."

Yeah, I thought I was going to be famous.

"How'd you realize you were wrong?"

After an embarrassingly long interval, it occurred to me to actually try *applying* Cantor's Diagonal Argument to my own construction. Since 1 is in {1} and 2 is in {2}, they wouldn't be in the resulting set, but 3, 4, 5 and everything else would be. And of course my construct didn't have the set {3, 4, 5, ...} anywhere in it. I'd mapped all the *finite* sets of integers onto integers, but none of the infinite sets.

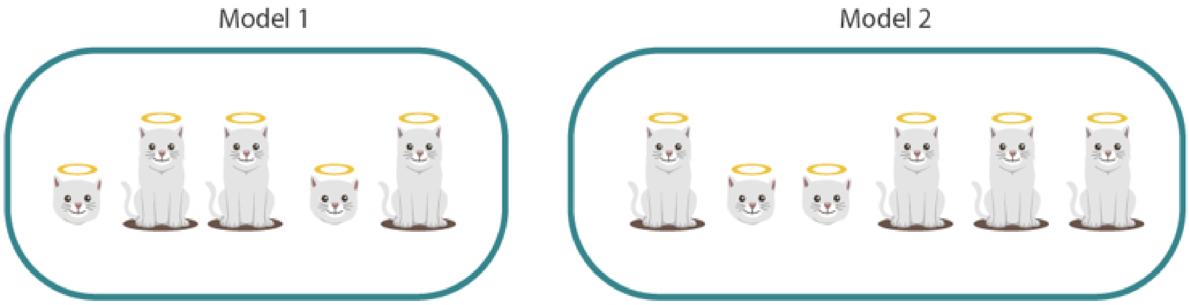
"Indeed."

I was then tempted to *go on* arguing that Cantor's Diagonal Argument was wrong *anyhow* because it was wrong to have infinite sets of integers. Thankfully, despite my young age, I was self-aware enough to realize I was being tempted to become a mathematical crank - I had also read a book on mathematical cranks by this point - and so I just quietly gave up, which was a valuable life lesson.

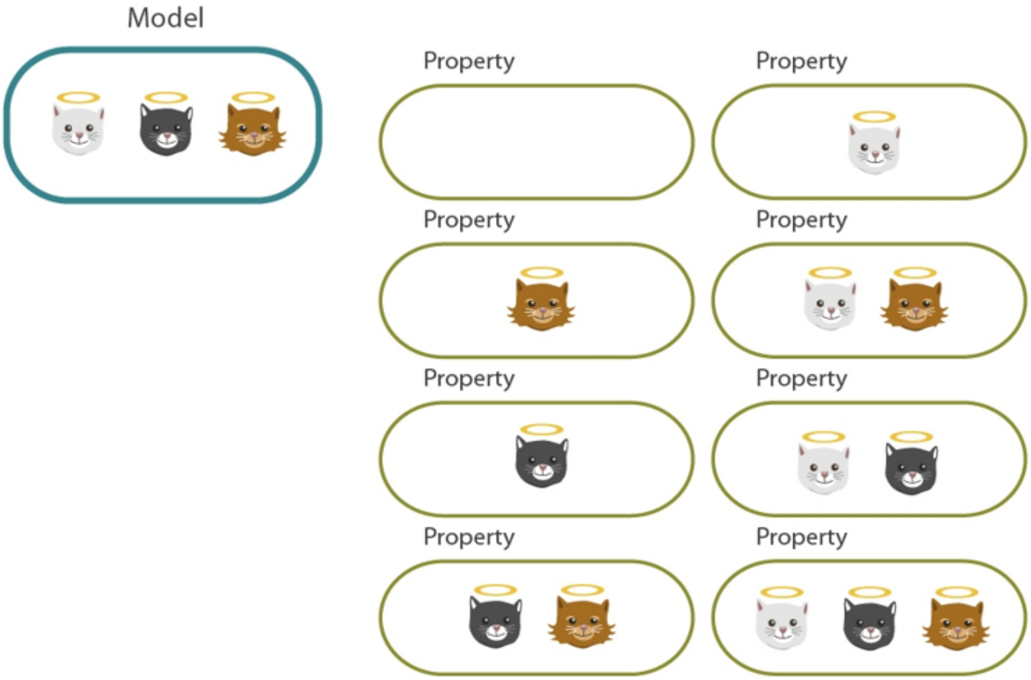
"Indeed."

But how exactly does Cantor's Diagonal Argument depend on second-order logic? Is it something to do with nonstandard integers?

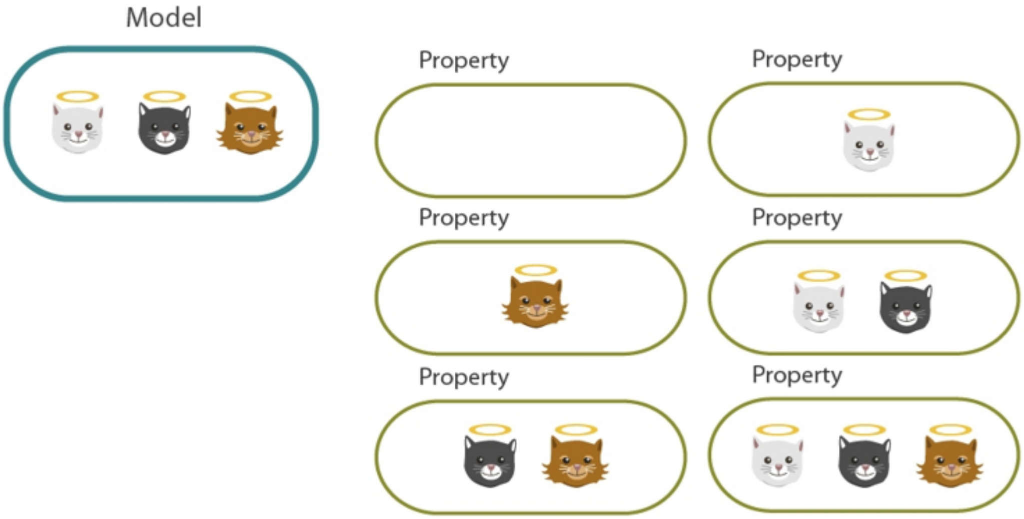
"Not exactly. What happens is that there's no way to make a first-order theory contain *all* subsets of an infinite set; there's no way to talk about *the* powerset of the integers. Let's illustrate using a finite metaphor. Suppose you have the axiom "All kittens are innocent." One model of that axiom might contain five kittens, another model might contain six kittens."



"In a second-order logic, you can talk about *all* possible collections of kittens - in fact, it's built into the syntax of the language when you quantify over all properties."



"In a first-order set theory, there are *some* subsets of kittens whose existence is provable, but others might be missing."



"Though that image is only metaphorical, since you *can* prove the existence of all the finite subsets. Just imagine that's an infinite number of kittens we're talking about up there."

And there's no way to say that *all possible* subsets exist?

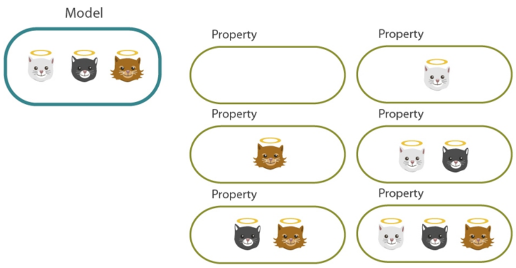
"Not in first-order logic, just like there's no way to say that you want as few natural numbers as possible. Let's look at it from the standpoint of first-order set theory. The Axiom of Powerset says:"

$$\forall A \, \exists P \, \forall B \, [B \in P \iff \forall C \, (C \in B \Rightarrow C \in A)]$$

Okay, so that says, for every set A, there exists a set P which is the *power set* of all subsets of A, so that for every set B, B is inside the powerset P *if and only if* every element of B is an element of A. Any set which contains only elements from A, will be inside the powerset of A. Right?

"Almost. There's just one thing wrong in that explanation - the word 'all' when you say 'all subsets'. The Powerset Axiom says that for any collection of elements from A, *if a set B happens to exist* which embodies that collection, that set B is inside the powerset P of A. There's no way of saying, within a first-order logical theory, that a set exists for *every possible* collection of A's elements. There may be *some* sub-collections of A whose existence you can prove. But other sub-collections of A will happen to exist as sets inside some models, but not exist in others."

So in the same way that first-order Peano arithmetic suffers from mysterious extra numbers, first-order set theory suffers from mysterious missing subsets.



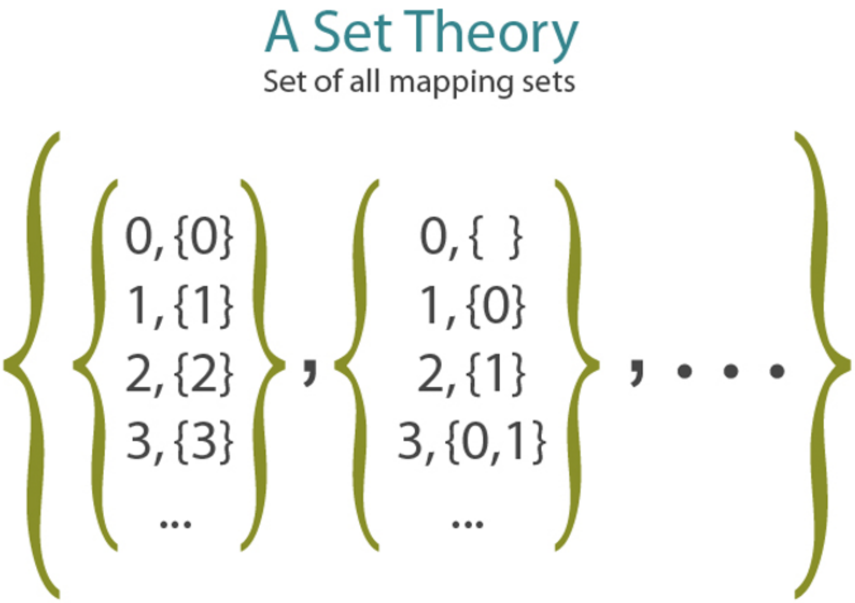
"Precisely. A first-order set theory might happen to be missing the particular infinite set corresponding to, oh, say, {3, 8, 17, 22, 28, ...} where the '...' is an infinite list of random numbers with no *compact* way of specifying them. If there's a compact way of specifying a set - if there's a finite formula that describes it - you can often prove it exists. But *most* infinite sets won't have any finite specification. It's precisely the claim to generalize over *all possible collections* that characterizes second-order logic. So it's trivial to say in a second-order set theory that *all* subsets exist. You would just say that for any set A, for any possible predicate P, there exists a set B which contains x iff x in A and Px."

I guess that torpedoes my clever idea about using first-order set theory to uniquely characterize the standard numbers by first asserting that there exists a set containing *at least* the standard numbers, and then talking about the *smallest subset* which obeys the Peano axioms.

"Right. When you talk about the numbers using first-order set theory, if there are *extra* numbers inside your set of numbers, the subset containing *just* the standard numbers must be missing from the powerset of that set. Otherwise you could find the smallest subset inside the powerset such that it contained 0 and contained the successor of every number it contained."

Hm. So then what exactly goes wrong with Cantor's Diagonal Argument?

"Cantor's Diagonal Argument uses the idea of a mapping between integers and sets of integers. In set theory, each mapping would itself be a set - in fact there would be a set of all mapping sets:"



"There's no way to first-order assert the existence of *every possible mapping* that *we* can imagine from outside. So a first-order version of the Diagonal Argument would show that in any *particular* model, for any mapping *that existed in the model* from integers to sets of integers, the model would also contain a diagonalized set of integers that wasn't in that mapping. This doesn't mean that *we* couldn't count all the sets of integers which *existed in the model.* The model could have so many 'missing' sets of integers that the remaining sets were denumerable. But then some mappings from integers to sets would also be missing, and in particular, the 'complete' mapping we can imagine from outside would be missing. And for every mapping that *was* in the model, the Diagonal Argument would construct a set of integers that wasn't in the mapping. On the outside, *we* would see a possible mapping from integers to sets - but that mapping wouldn't exist *inside* the model as a set. It takes a logic-of-collections to say that *all possible* integer-collections exist as sets, or that *no possible* mapping exists from the integers onto those sets."

So if first-order logic can't even talk about *finiteness* vs. *infiniteness* - let alone prove that there are *really* more sets of integers than integers - then why is anyone interested in first-order logic in the first place? Isn't that like trying to eat dinner using only a fork, when there are lots of interesting foods which *provably* can't be eaten with a fork, and you have a spoon?

"Ah, well... some people believe there *is* no spoon. But let's take that up next time."

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

Next post: "Second-Order Logic: The Controversy"

Previous post: "Standard and Nonstandard Numbers"

Participation in the LW Community Associated with Less Bias

Summary

CFAR included 5 questions on the 2012 LW Survey which were adapted from the heuristics and biases literature, based on five different cognitive biases or reasoning errors. LWers, on the whole, showed less bias than is typical in the published research (on all 4 questions where this was testable), but did show clear evidence of bias on 2-3 of those 4 questions. Further, those with closer ties to the LW community (e.g., those who had read more of the sequences) showed significantly less bias than those with weaker ties (on 3 out of 4-5 questions where that was testable). These results all held when controlling for measures of intelligence.

METHOD & RESULTS

Being less susceptible to cognitive biases or reasoning errors is one sign of rationality (see the work of Keith Stanovich & his colleagues, for example). You'd hope that a community dedicated to rationality would be less prone to these biases, so I selected 5 cognitive biases and reasoning errors from the heuristics & biases literature to include on the LW survey. There are two possible patterns of results which would point in this direction:

- high scores: LWers show less bias than other populations that have answered these questions (like students at top universities)
- correlation with strength of LW exposure: those who have read the sequences (or have been around LW a long time, have high karma, attend meetups, make posts) score better than those who have not.

The 5 biases were selected in part because they can be tested with everyone answering the same questions; I also preferred biases that haven't been discussed in detail on LW. On some questions there is a definitive wrong answer and on others there is reason to believe that a bias will tend to lead people towards one answer (so that, even though there might be good reasons for a person to choose that answer, in the aggregate it is evidence of bias if more people choose that answer).

This is only one quick, rough survey. If the results are as predicted, that could be because LW makes people more rational, or because LW makes people more familiar with the heuristics & biases literature (including how to avoid falling for the standard tricks used to test for biases), or because the people who are attracted to LW are already unusually rational (or just unusually good at avoiding standard biases). Susceptibility to standard biases is just one angle on rationality. Etc.

Here are the question-by-question results, in brief. The next section contains the exact text of the questions, and more detailed explanations.

Question 1 was a disjunctive reasoning task, which had a definitive correct answer. Only 13% of undergraduates got the answer right in the published paper that I took it from. 46% of LWers got it right, which is much better but still a very high error rate. Accuracy was 58% for those high in LW exposure vs. 31% for those low in LW

exposure. So for this question, that's:
1. LWers biased: yes
2. LWers less biased than others: yes
3. Less bias with more LW exposure: yes

Question 2 was a temporal discounting question; in the original paper about half the subjects chose money-now (which reflects a very high discount rate). Only 8% of LWers did; that did not leave much room for differences among LWers (and there was only a weak & nonsignificant trend in the predicted direction). So for this question:
1. LWers biased: not really
2. LWers less biased than others: yes
3. Less bias with more LW exposure: n/a (or no)

Question 3 was about the law of large numbers. Only 22% got it right in Tversky & Kahneman's original paper. 84% of LWers did: 93% of those high in LW exposure, 75% of those low in LW exposure. So:
1. LWers biased: a bit
2. LWers less biased than others: yes
3. Less bias with more LW exposure: yes

Question 4 was based on the decoy effect aka asymmetric dominance aka attraction effect (but missing a control condition). I don't have numbers from the original study (and there is no correct answer) so I can't really answer 1 or 2 for this question, but there was a difference based on LW exposure: 57% vs. 44% selecting the less bias related answer.
1. LWers biased: n/a
2. LWers less biased than others: n/a
3. Less bias with more LW exposure: yes

Question 5 was an anchoring question. The original study found an effect (measured by slope) of 0.55 (though it was less transparent about the randomness of the anchor; transparent studies w. other questions have found effects around 0.3 on average). For LWers there was a significant anchoring effect but it was only 0.14 in magnitude, and it did not vary based on LW exposure (there was a weak & nonsignificant trend in the wrong direction).
1. LWers biased: yes
2. LWers less biased than others: yes
3. Less bias with more LW exposure: no

One thing you might wonder: how much of this is just intelligence? There were several questions on the survey about performance on IQ tests or SATs. Controlling for scores on those tests, all of the results about the effects of LW exposure held up nearly as strongly. Intelligence test scores were also predictive of lower bias, independent of LW exposure, and those two relationships were almost the same in magnitude. If we extrapolate the relationship between IQ scores and the 5 biases to someone with an IQ of 100 (on either of the 2 IQ measures), they are still less biased than the participants in the original study, which suggests that the "LWers less biased than others" effect is not based solely on IQ.

MORE DETAILED RESULTS

There were 5 questions related to strength of membership in the LW community which

I standardized and combined into a single composite measure of LW exposure (LW use, sequence reading, time in community, karma, meetup attendance); this was the main predictor variable I used (time per day on LW also seems related, but I found out while analyzing last year's survey that it doesn't hang together with the others or associate the same way with other variables). I analyzed the results using a continuous measure of LW exposure, but to simplify reporting, I'll give the results below by comparing those in the top third on this measure of LW exposure with those in the bottom third.

There were 5 intelligence-related measures which I combined into a single composite measure of Intelligence (SAT out of 2400, SAT out of 1600, ACT, previously-tested IQ, extra credit IQ test); I used this to control for intelligence and to compare the effects of LW exposure with the effects of Intelligence (for the latter, I did a similar split into thirds). Sample sizes: 1101 people answered at least one of the CFAR questions; 1099 of those answered at least one LW exposure question and 835 of those answered at least one of the Intelligence questions. Further details about method available on request.

Here are the results, question by question.

Question 1: *Jack is looking at Anne, but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person?*

- *Yes*
- *No*
- *Cannot be determined*

This is a "disjunctive reasoning" question, which means that getting the correct answer requires using "or". That is, it requires considering multiple scenarios. In this case, either Anne is married or Anne is unmarried. If Anne is married then married Anne is looking at unmarried George; if Anne is unmarried then married Jack is looking at unmarried Anne. So the correct answer is "yes". A study by Toplak & Stanovich (2002) of students at a large Canadian university found that only 13% correctly answered "yes" while 86% answered "cannot be determined" (2% answered "no").

On this LW survey, 46% of participants correctly answered "yes"; 54% chose "cannot be determined" (and 0.4% said"no"). Further, correct answers were much more common among those high in LW exposure: 58% of those in the top third of LW exposure answered "yes", vs. only 31% of those in the bottom third. The effect remains nearly as big after controlling for Intelligence (the gap between the top third and the bottom third shrinks from 27% to 24% when Intelligence is included as a covariate). The effect of LW exposure is very close in magnitude to the effect of Intelligence; 60% of those in the top third in Intelligence answered correctly vs. 37% of those in the bottom third.

original study: 13%
weakly-tied LWers: 31%
strongly-tied LWers: 58%

Question 2: *Would you prefer to receive $55 today or $75 in 60 days?*

This is a temporal discounting question. Preferring $55 today implies an extremely (and, for most people, implausibly) high discount rate, is often indicative of a pattern of discounting that involves preference reversals, and is correlated with other biases. The question was used in a study by Kirby (2009) of undergraduates at Williams College (with a delay of 61 days instead of 60; I took it from a secondary source that said "60" without checking the original), and based on the graph of parameter values in that paper it looks like just under half of participants chose the larger later option of $75 in 61 days.

LW survey participants almost uniformly showed a low discount rate: 92% chose $75 in 61 days. This is near ceiling, which didn't leave much room for differences among LWers. For LW exposure, top third vs. bottom third was 93% vs. 90%, and this relationship was not statistically significant (p=.15); for Intelligence it was 96% vs. 91% and the relationship was statistically significant (p=.007). (EDITED: I originally described the Intelligence result as nonsignificant.)

original study: ~47%
weakly-tied LWers: 90%
strongly-tied LWers: 93%

Question 3: *A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day. In the smaller one, about 15 babies are born each day. Although the overall proportion of girls is about 50%, the actual proportion at either hospital may be greater or less on any day. At the end of a year, which hospital will have the greater number of days on which more than 60% of the babies born were girls?*

- *The larger hospital*
- *The smaller hospital*
- *Neither - the number of these days will be about the same*

This is a statistical reasoning question, which requires applying the law of large numbers. In Tversky & Kahneman's (1974) original paper, only 22% of participants correctly chose the smaller hospital; 57% said "about the same" and 22% chose the larger hospital.

On the LW survey, 84% of people correctly chose the smaller hospital; 15% said "about the same" and only 1% chose the larger hospital. Further, this was strongly correlated with strength of LW exposure: 93% of those in the top third answered correctly vs. 75% of those in the bottom third. As with #1, controlling for Intelligence barely changed this gap (shrinking it from 18% to 16%), and the measure of Intelligence produced a similarly sized gap: 90% for the top third vs. 79% for the bottom third.

original study: 22%
weakly-tied LWers: 75%
strongly-tied LWers: 93%

Question 4: *Imagine that you are a doctor, and one of your patients suffers from migraine headaches that last about 3 hours and involve intense pain, nausea, dizziness, and hyper-sensitivity to bright lights and loud noises. The patient usually needs to lie quietly in a dark room until the headache passes. This patient has a migraine headache about 100 times each year. You are considering three medications that you could prescribe for this patient. The medications have similar side effects,* *but differ in effectiveness and cost. The patient has a low income and must pay the cost because her insurance plan does not cover any of these medications. Which medication would you be most likely to recommend?*

- *Drug A: reduces the number of headaches per year from 100 to 30. It costs $350 per year.*
- *Drug B: reduces the number of headaches per year from 100 to 50. It costs $100 per year.*
- *Drug C: reduces the number of headaches per year from 100 to 60. It costs $100 per year.*

This question is based on research on the decoy effect (aka "asymmetric dominance" or the "attraction effect"). Drug C is obviously worse than Drug B (it is strictly dominated by it) but it is not obviously worse than Drug A, which tends to make B look more attractive by comparison. This is normally tested by comparing responses to the three-option question with a control group that gets a two-option question (removing option C), but I cut a corner and only included the three-option question. The assumption is that more-biased people would make similar choices to unbiased people in the two-option question, and would be more likely to choose Drug B on the three-option question. The model behind that assumption is that there are various reasons for choosing Drug A and Drug B; the three-option question gives biased people one more reason to choose Drug B but other than that the reasons are the same (on average) for more-biased people and unbiased people (and for the three-option question and the two-option question).

Based on the discussion on the original survey thread, this assumption might not be correct. Cost-benefit reasoning seems to favor Drug A (and those with more LW exposure or higher intelligence might be more likely to run the numbers). Part of the problem is that I didn't update the costs for inflation - the original problem appears to be from 1995 which means that the real price difference was over 1.5 times as big then.

I don't know the results from the original study; I found this particular example online (and edited it heavily for length) with a reference to Chapman & Malik (1995), but after looking for that paper I see that it's listed on Chapman's CV as only a "published abstract".

49% of LWers chose Drug A (the one that is more likely for unbiased reasoners), vs. 50% for Drug B (which benefits from the decoy effect) and 1% for Drug C (the decoy). There was a strong effect of LW exposure: 57% of those in the top third chose Drug A vs. only 44% of those in the bottom third. Again, this gap remained nearly the same when controlling for Intelligence (shrinking from 14% to 13%), and differences in Intelligence were associated with a similarly sized effect: 59% for the top third vs. 44% for the bottom third.

original study: ??
weakly-tied LWers: 44%
strongly-tied LWers: 57%

Question 5: *Get a random three digit number (000-999) from http://goo.gl/x45un and enter the number here.*

Treat the three digit number that you just wrote down as a length, in feet. Is the *height of the tallest redwood tree in the world more or less than the number that you wrote down?*

What is your best guess about the height of the tallest redwood tree in the world (in feet)?

This is an anchoring question; if there are anchoring effects then people's responses will be positively correlated with the random number they were given (and a regression analysis can estimate the size of the effect to compare with published results, which used two groups instead of a random number).

Asking a question with the answer in feet was a mistake which generated a great deal of controversy and discussion. Dealing with unfamiliar units could interfere with answers in various ways so the safest approach is to look at only the US respondents; I'll also see if there are interaction effects based on country.

The question is from a paper by Jacowitz & Kahneman (1995), who provided anchors of 180 ft. and 1200 ft. to two groups and found mean estimates of 282 ft. and 844 ft., respectively. One natural way of expressing the strength of an anchoring effect is as a slope (change in estimates divided by change in anchor values), which in this case is 562/1020 = 0.55. However, that study did not explicitly lead participants through the randomization process like the LW survey did. The classic Tversky & Kahneman (1974) anchoring question did use an explicit randomization procedure (spinning a wheel of fortune; though it was actually rigged to create two groups) and found a slope of 0.36. Similarly, several studies by Ariely & colleagues (2003) which used the participant's Social Security number to explicitly randomize the anchor value found slopes averaging about 0.28.

There was a significant anchoring effect among US LWers (n=578), but it was much weaker, with a slope of only 0.14 (p=.0025). That means that getting a random number that is 100 higher led to estimates that were 14 ft. higher, on average. LW exposure did not moderate this effect (p=.88); looking at the pattern of results, if anything the anchoring effect was slightly higher among the top third (slope of 0.17) than among the bottom third (slope of 0.09). Intelligence did not moderate the results either (slope of 0.12 for both the top third and bottom third). It's not relevant to this analysis, but in case you're curious, the median estimate was 350 ft. and the actual answer is 379.3 ft. (115.6 meters).

Among non-US LWers (n=397), the anchoring effect was slightly smaller in magnitude compared with US LWers (slope of 0.08), and not significantly different from the US LWers or from zero.

original study: slope of 0.55 (0.36 and 0.28 in similar studies)
weakly-tied LWers: slope of 0.09
strongly-tied LWers: slope of 0.17

If we break the LW exposure variable down into its 5 components, every one of the five is strongly predictive of lower susceptibility to bias. We can combine the first four CFAR questions into a composite measure of unbiasedness, by taking the percentage of questions on which a person gave the "correct" answer (the answer suggestive of lower bias). Each component of LW exposure is correlated with lower bias on that measure, with r ranging from 0.18 (meetup attendance) to 0.23 (LW use), all p < .0001 (time per day on LW is uncorrelated with unbiasedness, r=0.03, p=.39). For the composite LW exposure variable the correlation is 0.28; another way to express this relationship is that people one standard deviation above average on LW exposure 75% of CFAR questions "correct" while those one standard deviation below average got 61% "correct". Alternatively, focusing on sequence-reading, the accuracy rates were:

75%	Nearly all of the Sequences (n = 302)
70%	About 75% of the Sequences (n = 186)
67%	About 50% of the Sequences (n = 156)
64%	About 25% of the Sequences (n = 137)
64%	Some, but less than 25% (n = 210)
62%	Know they existed, but never looked at them (n = 19)
57%	Never even knew they existed until this moment (n = 89)

Another way to summarize is that, on 4 of the 5 questions (all but question 4 on the decoy effect) we can make comparisons to the results of previous research, and in all 4 cases LWers were much less susceptible to the bias or reasoning error. On 1 of the 5 questions (question 2 on temporal discounting) there was a ceiling effect which made it extremely difficult to find differences within LWers; on 3 of the other 4 LWers with a strong connection to the LW community were much less susceptible to the bias or reasoning error than those with weaker ties.

REFERENCES

Ariely, Loewenstein, & Prelec (2003), "Coherent Arbitrariness: Stable demand curves without stable preferences"
Chapman & Malik (1995), "The attraction effect in prescribing decisions and consumer choice"
Jacowitz & Kahneman (1995), "Measures of Anchoring in Estimation Tasks"
Kirby (2009), "One-year temporal stability of delay-discount rates"
Toplak & Stanovich (2002), "The Domain Specificity and Generality of Disjunctive Reasoning: Searching for a Generalizable Critical Thinking Skill"
Tversky & Kahneman's (1974), "Judgment under Uncertainty: Heuristics and Biases"

UFAI cannot be the Great Filter

[Summary: The fact we do not observe (and have not been wiped out by) an UFAI suggests the main component of the 'great filter' cannot be civilizations like ours being wiped out by UFAI. Gentle introduction (assuming no knowledge) and links to much better discussion below.]

Introduction

The Great Filter is the idea that although there is lots of matter, we observe no "expanding, lasting life", like space-faring intelligences. So there is some filter through which almost all matter gets stuck before becoming expanding, lasting life. One question for those interested in the future of humankind is whether we have already 'passed' the bulk of the filter, or does it still lie ahead? For example, is it very unlikely matter will be able to form self-replicating units, but once it clears that hurdle becoming intelligent and going across the stars is highly likely; or is it getting to a humankind level of development is not that unlikely, but very few of those civilizations progress to expanding across the stars. If the latter, that motivates a concern for working out what the forthcoming filter(s) are, and trying to get past them.

One concern is that advancing technology gives the possibility of civilizations wiping themselves out, and it is this that is the main component of the Great Filter - one we are going to be approaching soon. There are several candidates for which technology will be an existential threat (nanotechnology/'Grey goo', nuclear holocaust, runaway climate change), but one that looms large is Artificial intelligence (AI), and trying to understand and mitigate the existential threat from AI is the main role of the Singularity Institute, and I guess Luke, Eliezer (and lots of folks on LW) consider AI the main existential threat.

The concern with AI is something like this:

1. AI will soon greatly surpass us in intelligence in all domains.
2. If this happens, AI will rapidly supplant humans as the dominant force on planet earth.
3. Almost all AIs, even ones we create with the intent to be benevolent, will probably be unfriendly to human flourishing.

Or, as summarized by Luke:

> *... AI leads to intelligence explosion, and, because we don’t know how to give an AI benevolent goals, by default an intelligence explosion will optimize the world for accidentally disastrous ends. A controlled intelligence explosion, on the other hand, could optimize the world for good. (More on this option in the next post.)*

So, the aim of the game needs to be trying to work out how to control the future intelligence explosion so the vastly smarter-than-human AIs are 'friendly' (FAI) and make the world better for us, rather than unfriendly AIs (UFAI) which end up optimizing the world for something that sucks.

'Where is everybody?'

So, topic. I read this post by Robin Hanson which had a really good parenthetical remark (emphasis mine):

> *Yes, it is possible that the extremely difficultly was life's origin, or some early step, so that, other than here on Earth, all life in the universe is stuck before this early extremely hard step. But even if you find this the most likely outcome, surely given our ignorance you must also place a non-trivial probability on other possibilities. You must see a great filter as lying between initial planets and expanding civilizations, and wonder how far along that filter we are. In particular, you must estimate a substantial chance of "disaster", i.e., something destroying our ability or inclination to make a visible use of the vast resources we see.* ***(And this disaster can't be an unfriendly super-AI, because that should be visible.)***

This made me realize an UFAI should also be counted as an 'expanding lasting life', and should be deemed unlikely by the Great Filter.

Another way of looking at it: *if* the Great Filter still lies ahead of us, *and* a major component of this forthcoming filter is the threat from UFAI, we should expect to see the UFAIs of other civilizations spreading across the universe (or not see anything at all, because they would wipe us out to optimize for their unfriendly ends). That we do not observe it disconfirms this conjunction.

[**Edit/Elaboration**: It also gives a stronger argument - as the UFAI is the 'expanding life' we do not see, the beliefs, 'the Great Filter lies ahead' and 'UFAI is a major existential risk' lie opposed to one another: the higher your credence in the filter being ahead, the lower your credence should be in UFAI being a major existential risk (as the many civilizations like ours that go on to get caught in the filter do not produce expanding UFAIs, so expanding UFAI cannot be the main x-risk); conversely, if you are confident that UFAI is the main existential risk, then you should think the bulk of the filter is behind us (as we don't see any UFAIs, there cannot be many civilizations like ours in the first place, as we are quite likely to realize an expanding UFAI).]

A much more in-depth article and comments (both highly recommended) was made by Katja Grace a couple of years ago. I can't seem to find a similar discussion on here (feel free to downvote and link in the comments if I missed it), which surprises me: I'm not bright enough to figure out the anthropics, and obviously one may hold AI to be a big deal for other-than-Great-Filter reasons (maybe a given planet has a 1 in a googol chance of getting to intelligent life, but intelligent life 'merely' has a 1 in 10 chance of successfully navigating an intelligence explosion), but this would seem to be substantial evidence driving down the proportion of x-risk we should attribute to AI.

What do you guys think?

Beware Selective Nihilism

In a previous post, I argued that nihilism is often short changed around here. However I'm far from certain that it is correct, and in the mean time I think we should be careful not to discard our values one at a time by engaging in "selective nihilism" when faced with an ontological crisis, without even realizing that's what's happening. Karl recently reminded me of the post Timeless Identity by Eliezer Yudkowsky, which I noticed seems to be an instance of this.

As I mentioned in the previous post, our values seem to be defined in terms of a world model where people exist as ontologically primitive entities ruled heuristically by (mostly intuitive understandings of) physics and psychology. In this kind of decision system, both identity-as-physical-continuity and identity-as-psychological-continuity make perfect sense as possible values, and it seems humans do "natively" have both values. A typical human being is both reluctant to step into a teleporter that works by destructive scanning, and unwilling to let their physical structure be continuously modified into a psychologically very different being.

If faced with the knowledge that physical continuity doesn't exist in the real world at the level of fundamental physics, one might conclude that it's crazy to continue to value it, and this is what Eliezer's post argued. But if we apply this reasoning in a non-selective fashion, wouldn't we also conclude that we should stop valuing things like "pain" and "happiness" which also do not seem to exist at the level of fundamental physics?

In our current environment, there is widespread agreement among humans as to which macroscopic objects at time t+1 are physical continuations of which macroscopic objects existing at time t. We may not fully understand what exactly it is we're doing when judging such physical continuity, and the agreement tends to break down when we start talking about more exotic situations, and if/when we do fully understand our criteria for judging physical continuity it's unlikely to have a simple definition in terms of fundamental physics, but all of this is true for "pain" and "happiness" as well.

I suggest we keep all of our (potential/apparent) values intact until we have a better handle on how we're supposed to deal with ontological crises in general. If we convince ourselves that we should discard some value, and that turns out to be wrong, the error may be unrecoverable once we've lived with it long enough.

Narrative, self-image, and self-communication

Related to: Cached selves, Why you're stuck in a narrative, The curse of identity

Outline: Some back-story, Pondering the mechanics of self-image, The role of narrative, Narrative as a medium for self-communication.

tl;dr: One can have a self-image that causes one to neglect the effects of self-image. And, since we tend to process our self-images somewhat in the context of a narrative identity, if you currently make zero use of narrative in understanding and affecting how you think about yourself, it may be worth adjusting upward. All this seems to have been the case for me, and is probably part of what makes HPMOR valuable.

Some back-story

Starting when I was around 16 and becoming acutely annoyed with essentialism, I prided myself on not being dependent on a story-like image of myself. In fact, to make sure I wasn't, I put a break command in my narrative loop: I drafted a story in my mind about a hero who was able to outwit his foes by being less constrained by narrative than they were, and I identified with him whenever I felt a need-for-narrative coming on. Batman's narrator goes for something like this in the Dark Knight when he <select for spoiler->

I think this break command was mostly a good thing. It helped me to resolve cognitive dissonance and overcome the limitations of various cached selves, and I ended up mostly focussed on whether my beliefs were accurate and my desires were being fulfilled. So I still figure it's a decent first-order correction to being over-constrained by narrative.

But, I no longer think it's the only decent solution. In fact, understanding the more subtle mechanics of self-image — what affects our self schemas, what they affect, and how — was something I neglected for a long time because I saw self-image as a solved problem. Yes, I developed a cached view of myself as unaffected by self-image constraints. I would have been embarassed to notice such dependencies, so I didn't. The irony, eh?

I'm writing this because I wouldn't be surprised to find others here developing, or having developed, this blind spot...

Pondering the mechanics of self-image

At some point in your life, you may have taken on a job or a project without knowing that after doing it for a month, it would negatively affect your self-image in some way. There may have been things that you always found very easy to do which, after some aspect of your self-image changed, you suddenly found yourself avoiding or struggling with.

It would be nice to be able to predict and maybe even control that sort of thing in advance. In general, I'd like a deeper understanding of the following questions:

1. What actions might conflict or resonate with my self-image?
2. What events beyond my control might threaten or reinforce my self-image?
3. What might my self-image inhibit me from doing, or empower me to do?
4. Could changing my self-image help me further my goals?

If you've never sat to ask yourself these questions genuinely, I might suggest stopping here and thinking about them for a while. Simply taking the time to ponder these issues has lead me to many helpful realizations. For example:

- I used to be uninterested in how self-image worked because I didn't see myself as the kind of person who was affected by self-image!
- I didn't like dancing until I was 22, when I found a way to view it as a function of my "musician" self-schema.
- There were certain things I didn't try to learn about, like neuroscience, just because they didn't fit with my status-quo self-image as a mathematician. I noticed this acutely when I was was 23, after reading Anna's Cached Selves post, and I began reading a textbook on affective neuroscience.
- An injury that prevented me from climbing this semester lead to me feeling chronically *meh* for about a month, until I realized it was because my self-image as a physically active and playful person was threatened. Realizing this, and reconstructing my self-image as more generally "health-conscious", was how I got over it.

I don't have anything like an inclusive, general theory of self-image, and I have lots of hanging questions. Can I come up with a reasonably finite exhaustive list of features to track in my own self-image, for practical gains? Does such a list exist for people in general? But even without these, asking myself the old 1-4 once in a while gives me something to think about.

The role of narrative

In my experience, personally and with others, the answers to questions 1-4 are not automatically transparent, even if we can find partial answers by asking them directly. So what other questions can we ask ourselves to understand our self-images?

It seems to be common lore that our self-images have something to do with narrative identity. I take this to mean that we process our self-images somewhat in terms of features and schemas that we also use to process common stories.

So, I've tried working through the following series of questions to get in touch with what aspects of my personal narrative cause me to experience shame, pride, indignation, and nurturance. I like to lay them all out like this to signal to myself what they're for and that I want to do them all:

- Questions to understand shame:
  - I feel sad or ashamed when ...
  - When I'm sad or ashamed, I see myself as ... and I see the world as ...
  - Some real or fictional people, stories, songs, or poems I can relate to when I'm sad or ashamed:
- Questions to understand pride:
  - I feel happy or proud when ...

    - When I'm happy or proud, I see myself as ... and I see the world as ...
    - Some real or fictional people, stories, songs, or poems I can relate to when I'm happy or proud:
- Questions to understand indignation:
    - I feel angry or indignant when ...
    - When I am angry or indignant, I see myself as ... and I see the world as ...
    - Some real or fictional people, stories, songs, or poems I can relate to when I feel shame or indignation:
- Questions to understand nurturance:
    - I feel caring or nurturing when ...
    - When I am caring or nurturing, I see myself as ... and I see the world as ...
    - Some real or fictional people, stories, songs, or poems I can relate to when I feel caring or nurturing:

Consequences. By asking myself these questions, I've come to some realizations that didn't result from asking myself the more direct questions 1-4. For example:

- *(Involving shame and pride)* Doing physiotherapy exercises made me feel ashamed of being weak. Visualizing the anime character Naruto training to recover from injuries made me stop experiencing the exercises as a "sign of weakness", and I became less physically uncomfortable while doing them.
- *(Involving indignation and nurturance)* Imagining my kind and inspiring 6th grade teacher speaking to me an indignant tone of voice seems wrong, and makes me think that feeling annoyed is not always a good way to help other people learn from their mistakes, because he was the teacher I felt I learned the most moral lessons from growing up. "Channeling" him makes me more curious about other peoples' motives and misunderstandings instead of feeling annoyed.
- *(Involving all four)* Explicitly imagining myself as an <insert animal here> helps me to avoid taking myself too seriously — in particular, getting caught up in shame, indignation, and unhelpful instances of pride — while still caring about myself.

Does anyone have similar experiences they'd like to share? Or very dissimilar experiences? Or questions I could add to this list? Or well-reproduced psych references? HPMOR references are also highly encouraged, especially since I still haven't read it, and in light of this post, I probably should!

Narrative as a medium for self-communication

Like any method of affecting oneself, narrative is something one can over-use. But I think I personally have been over-cautious about this, to the point of neglecting it as an option and ignoring it as an unconscious constraint. To the extent that I now use it, I think of it as a way of communicating with myself, not to be used for trickery or over-selling a point.

To draw an analogy, if you tell your 2-year-old child "You trigger in me feelings of paternal nurturance", while this may be true, it's not communication. Hugging the child is communication. It's a language she'll understand. In fact, it's probably how you should teach her what "nurturance" means. In particular, it's not a trick, and it's not over-selling.

Likewise, when I'm convinced enough that something is true — like *for once I should really try not feeling annoyed with a postmodernist to see if we can communicate* — and it's time to tell that to my limbic system with some conviction, maybe it's worth speaking a language my emotional brain understands a little better, and maybe sometimes that language is narrative. Maybe I'll write myself a poem about patience. Maybe I already have ;)

Rational subjects and rational practitioners

Half-closing my eyes and looking at the recent topic of morality from a distance, I am struck by the following trend.

In mathematics, there are no substantial controversies. (I am speaking of the present era in mathematics, since around the early 20th century. There were some before then, before it had been clearly worked out what was a proof and what was not.) There are few in physics, chemistry, molecular biology, astronomy. There are some but they are not the bulk of any of these subjects. Look at biology more generally, history, psychology, sociology, and controversy is a larger and larger part of the practice, in proportion to the distance of the subject from the possibility of reasonably conclusive experiments. Finally, politics and morality consist of nothing but controversy and always have done.

Curiously, participants in discussions of all of these subjects seem equally confident, regardless of the field's distance from experimental acquisition of reliable knowledge. What correlates with distance from objective knowledge is not uncertainty, but controversy. Across these fields (not necessarily within them), opinions are firmly held, independently of how well they can be supported. They are firmly defended and attacked in inverse proportion to that support. The less information there is about actual facts, the more scope there is for continuing the fight instead of changing one's mind. (So much for the Aumann agreement of Bayesian rationalists.)

Perhaps mathematicians and hard scientists are not more rational than others, but work in fields where it is easier to be rational. When they turn into crackpots outside their discipline, they were actually that irrational already, but have wandered into an area without safety rails.

2012 Winter Fundraiser for the Singularity Institute

Cross-posted here.

(The Singularity Institute maintains Less Wrong, with generous help from Trike Apps, and much of the core content is written by salaried SI staff members.)

Thanks to the generosity of several major donors,† every donation to the Singularity Institute made now until January 20t (deadline extended from the 5th) will be matched dollar-for-dollar, up to a total of $115,000! So please, **donate now!**

Now is your chance to **double your impact** while helping us raise up to $230,000 to help fund our research program.

(If you're unfamiliar with our mission, please see our press kit and read our short research summary: Reducing Long-Term Catastrophic Risks from Artificial Intelligence.)

Now that Singularity University has acquired the Singularity Summit, and SI's interests in rationality training are being developed by the now-separate CFAR, **the Singularity Institute is making a major transition**. Most of the money from the Summit acquisition is being placed in a separate fund for a Friendly AI team, and therefore does not support our daily operations or other programs.

For 12 years we've largely focused on movement-building — through the Singularity Summit, Less Wrong, and other programs. This work was needed to build up a community of support for our mission and a pool of potential researchers for our unique interdisciplinary work.

Now, the time has come to say "Mission Accomplished Well Enough to Pivot to Research." Our community of supporters is now large enough that qualified researchers are available for us to hire, if we can afford to hire them. Having published 30+ research papers and dozens more original research articles on Less Wrong, we certainly haven't neglected research. But **in 2013 we plan to pivot so that a much larger share of the funds we raise is spent on research**.

Accomplishments in 2012

- Held a one-week research workshop on one of the open problems in Friendly AI research, and got progress that participants estimate would be the equivalent of 1-3 papers if published. (Details forthcoming. The workshop participants were Eliezer Yudkowsky, Paul Christiano, Marcello Herreshoff, and Mihaly Barasz.)
- Produced our annual Singularity Summit in San Francisco. Speakers included Ray Kurzweil, Steven Pinker, Daniel Kahneman, Temple Grandin, Peter Norvig, and many others.
- Launched the new Center for Applied Rationality, which ran 5 workshops in 2012, including Rationality for Entrepreneurs and SPARC (for young math

geniuses), and also published one (early-version) smartphone app, The Credence Game.
- Launched the redesigned, updated, and reorganized Singularity.org website.
- Achieved most of the goals from our August 2011 strategic plan.
- 11 new research publications.
- Eliezer published the first 12 posts in his sequence Highly Advanced Epistemology 101 for Beginners, the precursor to his forthcoming sequence, *Open Problems in Friendly AI*.
- SI staff members published many other substantive articles on Less Wrong, including How to Purchase AI Risk Reduction, How to Run a Successful Less Wrong Meetup, a Solomonoff Induction tutorial, The Human's Hidden Utility Function (Maybe), How can I reduce existential risk from AI?, AI Risk and Opportunity: A Strategic Analysis, and Checklist of Rationality Habits.
- Launched our new volunteers platform, SingularityVolunteers.org.
- Hired two new researchers, Kaj Sotala and Alex Altair.
- Published our press kit to make journalists' lives easier.
- And of course *much* more.

Future Plans You Can Help Support

In the coming months, we plan to do the following:

- As part of Singularity University's acquisition of the Singularity Summit, we will be changing our name and launching a new website.
- Eliezer will publish his sequence *Open Problems in Friendly AI*.
- We will publish nicely-edited ebooks (Kindle, iBooks, and PDF) for many of our core materials, to make them more accessible: *The Sequences, 2006-2009*, *Facing the Singularity*, and *The Hanson-Yudkowsky AI Foom Debate*.
- We will publish several more research papers, including "Responses to Catastrophic AGI Risk: A Survey" and a short, technical introduction to timeless decision theory.
- We will set up the infrastructure required to host a productive Friendly AI team and try hard to recruit enough top-level math talent to launch it.

(Other projects are still being surveyed for likely cost and strategic impact.)

We appreciate your support for our high-impact work! Donate now, and seize a better than usual chance to move our work forward. Credit card transactions are securely processed using either PayPal or Google Checkout. If you have questions about donating, please contact Louie Helm at (510) 717-1477 or louie@intelligence.org.

† $115,000 of total matching funds has been provided by Edwin Evans, Mihaly Barasz, Rob Zahra, Alexei Andreev, Jeff Bone, Michael Blume, Guy Srinivasan, and Kevin Fischer.

I will mostly be traveling (for AGI-12) for the next 25 hours, but I will try to answer questions after that.