

Best of LessWrong: December 2013

1. [Personal examples of semantic stopsigns](#)
2. [Why CFAR?](#)
3. [The Statistician's Fallacy](#)
4. [A critique of effective altruism](#)
5. [Building Phenomenological Bridges](#)
6. [Review of Scott Adams' "How to Fail at Almost Everything and Still Win Big"](#)
7. [Meditation: a self-experiment](#)
8. [How the Grinch Ought to Have Stolen Christmas](#)
9. [Doubt, Science, and Magical Creatures - a Child's Perspective](#)

Best of LessWrong: December 2013

1. [Personal examples of semantic stopsigns](#)
2. [Why CFAR?](#)
3. [The Statistician's Fallacy](#)
4. [A critique of effective altruism](#)
5. [Building Phenomenological Bridges](#)
6. [Review of Scott Adams' "How to Fail at Almost Everything and Still Win Big"](#)
7. [Meditation: a self-experiment](#)
8. [How the Grinch Ought to Have Stolen Christmas](#)
9. [Doubt, Science, and Magical Creatures - a Child's Perspective](#)

Personal examples of semantic stopsigns

I think most of us are familiar with the common [semantic stopsigns](#) like "God", "just because", and "it's a tradition." However, I've recently been noticing more interesting ones that I haven't really seen discussed on LW. (Or it's also likely that I missed those discussion.)

The first one is "humans are stupid." I notice this one very often, in particular in LW and other rationalist communities. The obvious problem here is that humans are not *that* stupid. Often what might seem like sheer stupidity was caused by a rather reasonable chain of actions and events. And even if a person or a group of people *is* being stupid, it's very interesting to chase down the cause. That's how you end up discovering biases from scratch or finding a great opportunity.

The second semantic stopsign is "should." Hat tip to Michael Vassar for bringing this one up. If you and I have a discussing about how I eat too much chocolate, and I say, "You are right, I should eat less chocolate," the conversation will basically end there. But 99 times out of a 100 nothing will actually come out of it. I try to taboo the word "should" from my vocabulary, so instead I will say something like, "You are right, I will not purchase any chocolate this month." This is a concrete *actionable* statement.

What other semantic stopsigns have you noticed in yourself and others?

Why CFAR?

Summary: We outline the case for CFAR, including:

- [Our long-term goal](#);
- [Our plan, and our progress to date](#);
- [Our financials](#); and
- [How you can help](#).

CFAR is in the middle of our [annual matching fundraiser](#) right now. If you've been thinking of donating to CFAR, now is the best time to decide for probably at least half a year. Donations up to \$150,000 will be matched until January 31st; and Matt Wage, who is matching the last \$50,000 of donations, has vowed *not* to donate unless matched.^[1]

Our workshops are cash-flow positive, and subsidize our basic operations (you are not subsidizing workshop attendees). But we can't yet run workshops often enough to fully cover our core operations. We also need to do more formal experiments, and we want to create free and low-cost curriculum with far broader reach than the current workshops. Donations are needed to keep the lights on at CFAR, fund free programs like the [Summer Program on Applied Rationality and Cognition](#), and let us do new and interesting things in 2014 (see below, at length).^[2]

Our long-term goal

CFAR's long-term goal is to create people who can and will solve important problems -- whatever the important problems turn out to be.^[3]

We therefore aim to create a community with three key properties:

1. **Competence** -- The ability to get things done in the real world. For example, the ability to work hard, follow through on plans, push past your fears, navigate social situations, organize teams of people, start and run successful businesses, [etc.](#)
2. **Epistemic rationality** -- The ability to form relatively accurate beliefs. Especially the ability to form such beliefs in cases where data is limited, motivated cognition is tempting, or the conventional wisdom is incorrect.
3. **Do-gooding** -- A desire to make the world better for all its people; the tendency to jump in and start/assist projects that might help (whether by labor or by donation); and ambition in keeping an eye out for projects that might help a lot and not just a little.

Why **competence**, **epistemic rationality**, and **do-gooding**?

To change the world, we'll need to be able to take effective action (competence).

We'll need to be able to form a good implicit and explicit understanding of the human world and how to shift it. We'll need to have the best shot we can get at modeling situations yet unseen. We'll need to solve problems outside the realms where competent business people already find traction (all of which require competence plus epistemic rationality). And we'll need to blend these abilities with a burning ambition

to leave the world far better than we found it (competence plus epistemic rationality plus do-gooding).

And we'll need a community, not just a set of individuals. It is hard for an isolated individual to figure out what the most important problems are, let alone how to effectively solve them. This is still harder for individuals who have interesting day jobs, and who are busy amassing real-world competence of varied sorts. Communities can assemble a complex world-model piece by piece. Communities can build and sustain motivation, as well, and facilitate the practice and transfer of useful skills. The aim is thus to create a community that, collectively, can figure out what needs doing and can then do it -- even when this requires multiple simultaneous competencies (e.g., locating a particular existential risk, *and* having good scientific connections, *and* knowing good folks in policy, *and* knowing how to do good technical research).

We intend to build that sort of community.

Our plan, and our progress to date

How can we create a community with high levels of competence, epistemic rationality, and do-gooding? By creating curricula that teach (or enhance) these properties; by seeding the community with diverse competencies and diverse perspectives on how to do good; and by linking people together into the right kind of community.

We've now had two years to execute on this vision.^[4] It's not a *lot* of time, but it's enough to get started; and it's enough that folks should already be able to update as to our ability to execute.

Here's our current working plan, the progress we've made so far, and the pieces we still need to hit.

Curriculum design

In October 2012, we had no money and little visible means of obtaining more.^[5] We needed runway; and we needed a way to use that runway to rapidly iterate curriculum.

We therefore focused our initial efforts into making a workshop that could pay its own bills, and at the same time give us data -- a workshop that would give us the opportunity to run (and learn from) many further workshops. Our [applied rationality workshops](#) have filled this role.

Progress to date

Reported benefits

After about a dozen workshops (and over 100 classes that we've designed and tested), we've settled on a workshop model that runs smoothly, and seems to provide value to our participants, who report a mean of 9.3 out of 10 to the question "Are you glad you came?". In the process we've substantially improved

our skill at curriculum design: it used to take us about 40 hours to design a unit we regarded as decent (design; test on volunteers; re-design; test on volunteers; etc). It now takes us about 8 hours to design a unit of the same quality.[\[6\]](#)

Anecdotally, we have many, many stories from alumni about how our workshop increased their competence (both generally and for altruistic ends). For example, alum Ben Toner, CEO of Draftable, recounts that after the July 2012 workshop, "At work, I realized I wasn't doing anywhere near enough planning. My employees were spending time on the wrong things because I hadn't planned things out in enough detail to make it clear what was the most important thing to do next. I fixed this immediately after the camp." Alum Ben Kuhn has described how the CFAR workshop helped his effective altruism group "vastly increase our campus presence--everything from making uncomfortable cold calls to powering through bureaucracy, and from running complex events to quickly updating on feedback." (Check out our [testimonials](#) page for more examples.)

Measurement

Anecdotal notwithstanding, the jury is still out regarding the workshops' usefulness to those who come. During the very first minicamps (the current workshops are agreed to be better) we randomized admission of 15 applicants, with 17 controls. Our study was low-powered and effects on e.g. income would have needed to be very large for us to expect to detect them. Still, we ended up with non-negligible [evidence of absence](#): income, happiness, and exercise did not visibly trend upward one year later. We detected statistically significant positive impacts on the standard ([BFI-10](#)) survey pair for emotional stability "I see myself as someone who is relaxed, handles stress well" / "I get nervous easily" ($p=.002$).

Also significant were effects on an abridged [General Self-Efficacy Scale](#) (sample item: "I can solve most problems if I invest the necessary effort") ($p=.007$). The details will be available soon on our blog (including a much larger number of negative results). We'll run another RCT soon, funding permitting.

Like many participants, we at CFAR have the subjective impression that the workshops boost strategicness; and, like most who have observed two workshops, we have the impression that today's workshops are much better than those in the initial RCT. We'll need to find ways to actually test those impressions, and to create stronger feedbacks from measurement into curriculum development.

Epistemic rationality curricula

After a rocky start, our epistemic rationality curriculum has seen a number of recent victories. Our "Building Bayesian Habits" class began performing much better after we figured out how to help people notice their intuitive, "System 1" expectations of probabilities.[\[7\]](#) Our "inner simulator" class conveys the distinction between profession and anticipation while aiming at immediate, practical benefits; it isn't about religion and politics, it's about whether your mother will actually enjoy the potted plant you're thinking of giving her. More generally, the epistemic rationality curriculum [appears to be](#) integrating deeply with the competence curriculum, and appears to be becoming more appealing to participants as it does so. Strengthening this curriculum, and building in real tests of its efficacy, will be a major focus in 2014.

Integrating with academic research

We made preliminary efforts in this direction - for example by taking standard questionnaires from the academic literature, including [Stanovich's](#) indicators of the traits he calls "rationality", and administering them to attendees at a Less

Wrong meetup. (We found that meetup attendees scored near the ceiling, so we'll probably need new questionnaires with better discrimination.) Our research fellow, Dan Keys (whose masters thesis was on heuristics and biases), spends a majority of his time keeping up with the literature and integrating it with CFAR workshops, as well as designing tests for our ongoing [forays](#) into randomized controlled trials. We're particularly excited by Tetlock's [Good Judgment Project](#), and we'll be piggybacking on it a bit to see if we can get decent ratings.

Accessibility

Initial workshops worked only for those who had already read the LW Sequences. Today, workshop participants who are smart and analytical, but with no prior exposure to rationality -- such as a local politician, a police officer, a Spanish teacher, and others -- are by and large quite happy with the workshop and feel it is valuable.

Nevertheless, the total set of people who can travel to a 4.5-day immersive workshop, and who can spend \$3900 to do so, is limited. We want to eventually give a substantial skill-boost in a less expensive, more accessible format; we are slowly bootstrapping toward this.

Specifically:

- **Shorter workshops:** We're working on shorter versions of our workshops (including [three-hour](#) and [one-day](#) courses) that can be given to larger sets of people at lower cost.
- **College courses:** We helped develop a [course on rational thinking](#) -- for UC Berkeley undergraduates, in partnership with Nobel Laureate Saul Perlmutter. We also brought several high school and university instructors to our workshop, to help seed early experimentation into their curricula.
- **Increasing visibility:** We've been working on increasing our visibility among the general public, with alumni James Miller and Tim Czech both working on non-fiction books that feature CFAR, and several mainstream media articles about CFAR on their way, including one forthcoming shortly in the Wall Street Journal.

Next steps

In 2014, we'll be devoting more resources to epistemic curriculum development; to research measuring the effects of our curriculum on both competence and epistemic rationality; and to more widely accessible curricula.

Forging community

The most powerful interventions are not one-off experiences; rather, they are the start of an ongoing practice. Changing one's social environment is [one of the highest impact ways to create personal change](#). Alum Paul Crowley writes that "The most valuable lasting thing I got out of attending, I think, is a renewed determination to continually up my game. A big part of that is that the minicamp creates a lasting community of fellow alumni who are also trying for the biggest bite of increased utility they can get, and that's no accident."

The goal is to create a community that is directly helpful for its members, and that simultaneously improves its members' impact on the world.

Progress to date

A strong set of seed alumni

We have roughly 350 alumni so far, which include scientists from MIT and Berkeley, college students, engineers from Google and Facebook, founders of Y-combinator startups, teachers, professional writers, and the exceptionally gifted high-school students who participated in [SPARC](#) 2013 and 2012. (Not counted in that tally are the 50-some attendees of the [2013 Effective Altruism Summit](#), for whom we ran a free, abridged version of our workshop.)

Alumni contact/community

There is an active alumni Google group, which gets daily traffic. Alumni use it to share useful life hacks they've discovered, help each other trouble-shoot, and notify each other of upcoming events and opportunities. We've also been using our post-workshop parties as reunions for alumni nearby (in the San Francisco Bay area, the New York City area, and -- in two months -- Melbourne, Australia).

In large part thanks to our alumni forum and the post-workshop party networking, there have already been numerous cases of alumni helping each other find jobs and collaborating on startups or other projects. There have also been several alumni recruited to do-gooding projects (e.g., MIRI and Leverage Research have engaged multiple alumni), and of alumni improving their "earn to give" ability or shifting their own do-gooding strategy.

Many alumni also take CFAR skills back to Less Wrong meet-ups or other local communities (for example, the effective-altruism meetup in Melbourne, a homeless youth shelter in Oregon, and a self-improvement group in NYC; many have also practiced in their start-ups and with co-workers (for example, Beeminder, MetaMed, and Aquahug)).

Do-gooding diversity

We'd like the alumni community to have an accurate picture of how to effectively improve the world. We don't want to try to figure out how to improve the world all from scratch. There are already a number of groups who've done a lot of good thinking on the subject; including some who call themselves "effective altruists", but also people who call themselves "social entrepreneurs", "x-risk minimizers", and "philanthropic foundations".

We aim to bring in the best thinkers and doers from all of these groups to seed the community with diverse good ideas on the subject. **The goal is to create a culture rich enough that the alumni, as a community, can overcome any errors in CFAR's founders' perspectives.** The goal is also to create a community that is defined by its pursuit of true beliefs, and that is not defined by any particular preconceptions as to what those beliefs are.

We use applicants' inclination to do good as a major criterion of financial aid. Recipients of our informally-dubbed "altruism scholarships" have included members of the Future of Humanity Institute, CEA, Giving What We Can, MIRI, and Leverage Research. They also include many college or graduate students who have no official EA affiliation, but who are passionate about their desire to devote their career to world-saving (and who hope the workshops can help them figure out how to do so). And they include folks who are working full-time on varied do-

gooding projects of broader origin, such as social entrepreneurs, someone working on community policing, and folks working at a major philanthropic foundation.

International outreach

We'll be running our first international workshop in Australia, in February 2014, thanks to alumni Matt and Andrew Fallshaw.

Also, starting in 2014, we'll be bringing about 20 Estonian math and science award-winners per year to CFAR workshops, thanks to a 5-year pledge from Jaan Tallinn to sponsor workshop spots for leading students from his home country.

Estonia is an EU member country with a population of 1.2 million and a high-technology economy, and going forward this might be the first opportunity to check whether there are network effects in relatively larger fractions of a stratum.

Next steps

Over 2014, a major focus will be improving opportunities for ongoing alumni involvement. If funding allows, we'll also try our hand at pilot activities for meet-ups.

Specific plans include:

- A two-day "Epistemic Rationality and EA" mini-workshop in January, targeted at alumni
- An alumni reunion this summer (which will be a multi-day event drawing folks our entire worldwide alumni community, unlike the alumni parties at each workshop);
- An alumni directory, as an attempt to increase business and philanthropic partnerships among alumni.

Financials

Expenses

Our fixed expenses come to about **\$40k per month**. In some detail:

- About \$7k for our office space
- About \$3k for miscellaneous expenses
- About \$30k for salary & wages, going forward
 - We have five full-time people on salary, each getting \$3.5k per month gross. The employer portion of taxes adds roughly an additional \$1k/month per employee.
 - The remaining \$7k or so goes to hourly employees and contractors. We have two roughly full-time hourly employees, and a few contractors who do website adjustment and maintenance, workbook compilation for a workshop, and similarly targeted tasks.

In addition to our fixed expenses, we chose to run [SPARC 2013](#), even though it would cause us to run out of money right around the end-of-year fundraising drive. We did so

because we judged SPARC to be potentially very important^[8], enough to justify the risk of leaning on this winter fundraiser to continue. All told, SPARC cost approximately \$50k in direct costs (not counting staff time).

(We also chose to e.g. teach at the EA Summit, do rationality research, put some effort into curricula that can be delivered cheaply to a larger crowd, etc. These did not incur much direct expense, but did require staff time which could otherwise have been directed towards revenue-producing projects.)

Revenue

Workshops are our primary source of non-donation income. We ran 7 of them in 2013, and they became increasingly cash-positive through the year. We now expect a full 4-day workshop held in the Bay Area to give us a profit of about \$25k (ignoring fixed costs, such as staff time and office rent), which is just under 3 weeks of CFAR runway.

Demand isn't yet reliable enough to let us run them at that frequency. We've made significant traction on building interest outside of the Less Wrong community, but there's still work to be done here, and that work will take time. In the meantime, workshops can subsidize some of our non-workshop activities, but not all of them. (Your donations do not go to subsidize workshops!)

We're also actively exploring revenue models other than the four-day workshop. Several of them look promising, but need time to come to fruition before the income they offer us is relevant.

Donations

CFAR received \$166k in our previous fundraising drive at the start of 2013, and a smaller amount of donations spread across the rest of the year. SPARC was partially sponsored with \$15k from Dropbox and \$5k from Quixey. These donations subsidized SPARC, the rationality workshop at the EA summit, research and development, and core expenses and salary.

Savings and debt

Right now CFAR has essentially no savings. The savings we accumulated by the end of 2012 went to (a) feeding the gap between income and expenses and (b) funding SPARC.

A \$30k loan, which helped us cover core 2013 expenses, comes due in March 2014.

Summary

If this winter fundraiser goes well, it will give us time to make some of our current experimental products mature. We think we have an excellent shot at making major

strides forward in CFAR's mission as well as becoming much more self-sustaining during 2014.

If this winter fundraiser goes poorly, CFAR will not yet have sufficient funding to continue core operations.

How you can help

Our main goals in 2014:

1. Building a **scalable revenue base**, including via ramping up our workshop quality, workshop variety, and our marketing reach.
2. **Community-building**, including an alumni reunion.
3. Creating more connections with the **effective altruism community**, and other opportunities for our alumni to get involved in do-gooding.
4. **Research** to feed back into our curriculum -- on the effectiveness of particular rationality techniques, as well as the long-term impact of rationality training on meaningful life outcomes.
5. Developing more classes on **epistemic rationality**.

The three most important ways you can help:

1. Donations

If you're considering donating but want to learn more about how CFAR uses money, or you have other questions or hesitations, let us know -- we'd be more than happy to chat with you via Skype. You can sign up for a one-on-one call with Anna [here](#).

2. Talent

We're actively seeking a new [director of operations](#) to organize our workshops; good operations can be a great multiplier on CFAR's total ability to get things done. We are continuing to try out exceptional candidates for a [curriculum designer](#).^[9] And we always need more [volunteers](#) to help out with alpha-testing new classes in Berkeley, and to participate in online experiments.

3. Participants

We're continually searching for additional awesome people for our workshops. This really is a high-impact way people can help us; and we do have a large amount of data suggesting that (you /your friends) will be glad to have come. You can apply [here](#) -- it takes 1 minute, and leads to a conversation with Anna or Kenzi, which (you'll / they'll) probably find interesting whether or not they choose to come.

Like the open-source movement, applied rationality will be the product of thousands of individuals' contributions. The ideas we've come up with so far are only a beginning. If you have other suggestions for people we should meet, other workshops we should attend, ways to branch out from our current business model, or anything else -- get in touch, we'd love to Skype with you.

You can also be a part of open-source applied rationality by creating good content for Less Wrong. Some of our best workshop participants, volunteers, hires, ideas for rationality techniques, use cases, and general inspiration have come from Less Wrong. Help keep the LW community vibrant and growing.

And, if you're willing -- do consider [donating now](#).

Footnotes

[1] That is: by giving up a dollar, you can, given some simplifications, cause CFAR to gain two dollars. Much thanks to Matt Wage, Peter McCluskey, Benjamin Hoffman, Janos Kramar & Victoria Krakovna, Liron Shapira, Satvik Beri, Kevin Harrington, Jonathan Weissman, and Ted Suzman for together putting up \$150k in matching funds. (Matt Wage, as mentioned, promises not only that he will donate if the pledge is matched, but also that he *won't* donate the \$50k of matching funds to CFAR if the pledge *isn't* filled -- so your donation probably really does cause matching at the margin.)

[2] This post was result of a collaborative effort between Anna Salamon, Kenzi Amodei, Julia Galef, and "Valentine" Michael Smith - like many of our endeavors at CFAR, it went through many iterations, in many hands, to create an overall whole where the credit due is difficult to tease apart.

[3] In the broadest sense, CFAR can be seen as a cognitive branch of effective altruism - making a marginal improvement to thinking where thinking matters a lot. MIRI did not gain traction until it began to include explicit rationality in its message - maybe because thinking about AI puts heavy loads on particular cognitive skills, though there are other hypotheses. Other branches of effective altruism may encounter their own problems with a heavy cognitive load. Effective altruism is limited in its growth by the supply of competent people who want to quantify the amount of good they do.

It has been true over the course of human history that improvements in world welfare have often been tied to improvements in explicit thinking skills, most notably with the invention of science. Even for someone who doesn't think that existential risk is the right place to look, trying to invest more in good reasoning, qua good reasoning - doubling down on the huge benefits which explicit cognitive skills have already brought humanity - is a plausible candidate for the highest-impact marginal altruism.

[4] That is, we've had two years since our barest beginnings, when Anna, Julia, and Val began working together under the auspices of MIRI; and just over a year as a financially and legally independent organization.

[5] Our pilot minicamps, prior to that October, gave us valuable data/iteration; but they did not pay for their own direct (room and board) costs, let alone for the staff time required.

[6] I'm estimating quality by workshop participants' feedback, here; it takes many fewer hours now for our instructors to create units that receive the same participant ratings as some older unit that hasn't been revised (we did this accidental experiment several times). Unsurprisingly, large quantities of unit-design practice, with rapid iteration and feedback, were key to improving our curriculum design skills.

[7] Interestingly, we threw away over a dozen versions of the Bayes class before we developed this one. It has proven somewhat easier to create curricula around strategicness, and around productivity/effectiveness more generally, than around epistemic rationality. The reason for the relative difficulty appears to be two-fold.

First, it is somewhat harder to create a felt need for epistemic rationality skills, at least among those who aren't working on gnarly, data-sparse problems such as existential risk. Second, there is more existing material on strategicness than on epistemic rationality; and it is in general harder to create from scratch than to create with borrowing. Nevertheless, we have, via much iteration, had some significant successes, including Bayes, separating professed beliefs from anticipated ones, and with certain subskills of avoiding motivated cognition (e.g. noticing curiosity; noticing and tuning in to mental flinches). Better yet, there seems to be a pattern to these successes which we are gradually getting the hang of.

We're excited that Ben Hoffman has pledged \$23k of funding specifically to enable us to improve our epistemic rationality curriculum and our research plan.

[8] From the perspective of long-term, high-impact altruism, highly math-talented people are especially worth impacting for a number of reasons. For one thing, if AI does turn out to pose significant risks over the coming century, there's a significant chance that at least one key figure in the eventual development of AI will have had amazing math tests in high school, judging from the history of past such achievements. An eventual scaled-up SPARC program, including math talent from all over the world, may be able to help that unknown future scientist build the competencies he or she will need to navigate that situation well.

More broadly, math talent may be relevant to other technological breakthroughs over the coming century; and tech shifts have historically impacted human well-being quite a lot relative to the political issues of any given day.

[9] To those who've already applied: Thanks very much for applying; and our apologies for not getting back to you so far. If the funding drive is filled (so that we can afford to possibly hire someone new), we'll be looking through the applications shortly after the drive completes and will get back to you then.

The Statistician's Fallacy

[**Epistemic status** | Contains generalization based on like three data points.]

In grad school, I took a philosophy of science class that was based around looking for examples of bad reasoning in the scientific literature. The kinds of objections to published scientific studies we talked about were not stupid ones. The professor had a background in statistics, and as far as I could tell knew her stuff in that area (though she dismissed Bayesianism in favor of frequentism). And no, unlike [some of the professors in the department](#), she wasn't an anti-evolutionist or anything like that.

Instead she was convinced that cellphones cause cancer. In spite of the fact that there's [scant evidence](#) for that claim, and there's [no plausible physical mechanism for how that could happen](#). This along with a number of other borderline-fringe beliefs that I won't get into here, but that was the big screaming red flag.*

Over the course of the semester, I got a pretty good idea of what was going on. She had an agenda—it happened to be an environmentalist, populist, pro-"natural"-things agenda, but that's incidental. The problem was that when she saw a scientific study that seemed at odds with her agenda, she went looking for flaws. And often she could find them! Real flaws, not ones she was imagining! But people who've read the [rationalization sequence](#) will see a problem here...

In my [last post](#), I quoted Robin Hanson on the tendency of some physicists to be unduly dismissive of other fields. But based the above case and a couple others like it, I've come to suspect statistics may be even *worse* than physics in that way. That fluency in statistics sometimes causes a supercharged [sophistication effect](#).

For example, some anthropogenic global warming skeptics make a big deal of alleged statistical errors in global warming research, but as I wrote in my post [Trusting Expert Consensus](#):

Michael Mann et al's so-called "[hockey stick](#)" graph has come under a lot of fire from skeptics, but (a) many other reconstructions have reached the same conclusion and (b) a panel formed by the National Research Council concluded that, while there were some problems with Mann et al's statistical analysis, these problems did not affect the conclusion. Furthermore, even if we didn't have the pre-1800 reconstructions, I understand that given what we know about CO2's heat-trapping properties, and given the increase in atmospheric CO2 levels due to burning fossil fuels, it would be surprising if humans *hadn't* caused significant warming.

Most recently, I got into a Twitter argument with someone who claimed that "IQ is demonstrably statistically meaningless" and that this was widely accepted among statisticians. Not only did this set off my ["academic clique!"](#) alarm bells, but I'd just come off doing a spurt of reading about intelligence, including the excellent [Intelligence: A Very Short Introduction](#). The claim that IQ is meaningless was wildly contrary to what I understood was the consensus among people who study intelligence for a living.

In response to my surprise, I got [an article](#) that contained lengthy and impressive-looking statistical arguments... but completely ignored a couple key points from the intelligence literature I'd read: first, that there's a strong correlation between IQ and

real-world performance, and second that correlations between the components of intelligence we know how to test for turn out to be really strong. If IQ is actually made up of several independent factors, we haven't been able to find them. Maybe some people in intelligence research really did make the mistakes alleged, but there was more to intelligence research than the statistician who wrote the article let on.

It would be fair to [shout a warning about correspondence bias](#) before inferring anything from these cases. But consider two facts:

1. Essentially all scientific fields rely heavily on statistics.
2. There's a lot more to mastering a scientific discipline than learning statistics, which limits how well most scientists will ever master statistics.

The first fact may make it tempting to think that if you know a lot of statistics, you're in a privileged position to judge the validity of any scientific claim you come across. But the second fact means that if you've *specialized* in statistics, you'll probably be better at it than most scientists, even good scientists. So if you go scrutinizing their papers, there's a good chance you'll find clear mistakes in their stats, and an even better chance you'll find arguable ones.

Bayesians will realize that, since there's a good chance that of happening *even when the conclusion is correct and well-supported by the evidence*, finding mistakes in the statistics is only weak evidence that the conclusion is wrong. Call it the statistician's fallacy: thinking that finding a mistake in the statistics is sufficient grounds to dismiss a finding.

Oh, if you're dealing with a novel finding that experts in the field aren't sure what to make of yet, and the statistics turns out to be wrong, then that may be enough. You may have better things to do than investigate further. But when [a solid majority of the experts](#) agree on a conclusion, and you see flaws in their statistics, I think the default assumption should be that they still know the issue better than you and very likely the sum total of the available evidence *does* support the conclusion. Even if the specific statistical arguments you've seen from them are wrong.

**Note: I've done some Googling to try to find rebuttals to this link, and most of what I found confirms it. I did find some people talking about multi-photon effects and heating, but couldn't find defenses of these suggestions that rise beyond people saying, ["well there's a chance."](#)*

A critique of effective altruism

I recently ran across Nick Bostrom's idea of subjecting your strongest beliefs to a [hypothetical apostasy](#) in which you try to muster the strongest arguments you can against them. As you might have figured out, I believe strongly in [effective altruism](#)—the idea of applying evidence and reason to finding the best ways to improve the world. As such, I thought it would be productive to write a hypothetical apostasy on the effective altruism movement.

(EDIT: As per the comments of [Vaniver](#), [Carl Shulman](#), and others, this didn't quite come out as a hypothetical apostasy. I originally wrote it with that in mind, but decided that a focus on more plausible, more moderate criticisms would be more productive.)

How to read this post

(EDIT: the following two paragraphs were written before I softened the tone of the piece. They're less relevant to the more moderate version that I actually published.)

Hopefully this is clear, but as a disclaimer: this piece is written in a fairly critical tone. This was part of an attempt to get "in character". *This tone does not indicate my current mental state with regard to the effective altruism movement.* I agree, to varying extents, with some of the critiques I present here, but I'm not about to give up on effective altruism or stop cooperating with the EA movement. The apostasy is purely hypothetical.

Also, because of the nature of a hypothetical apostasy, I'd guess that for effective altruist readers, the critical tone of this piece may be especially likely to trigger defensive rationalization. Please read through with this in mind. (A good way to counteract this effect might be, for instance, to imagine that you're not an effective altruist, but your friend is, and it's them reading through it: how should they update their beliefs?)

(End less relevant paragraphs.)

Finally, if you've never heard of effective altruism before, I don't recommend making this piece your first impression of it! You're going to get a very skewed view because I don't bother to mention all the things that are awesome about the EA movement.

Abstract

Effective altruism is, to my knowledge, the first time that a substantially useful set of ethics and frameworks to analyze one's effect on the world has gained a broad enough appeal to resemble a social movement. (I'd say these principles are something like *altruism*, *maximization*, *egalitarianism*, and *consequentialism*; together they imply many improvements over the social default for trying to do good in the world—earning to give as opposed to doing direct charity work, working in the developing world rather than locally, using evidence and feedback to analyze effectiveness, etc.) Unfortunately, as a movement effective altruism is failing to use these principles to acquire correct nontrivial beliefs about how to improve the world.

By way of clarification, consider a distinction between two senses of the word “trying” I used above. Let’s call them “actually trying” and “pretending to try”. Pretending to try to improve the world is something like responding to social pressure to improve the world by querying your brain for a thing which improves the world, taking the first search result and rolling with it. For example, for a while I thought that I would try to improve the world by developing computerized methods of checking informally-written proofs, thus allowing more scalable teaching of higher math, democratizing education, etc. *Coincidentally*, computer programming and higher math happened to be the two things that I was best at. This is pretending to try. Actually trying is looking at the things that improve the world, figuring out which one maximizes utility, and then doing that thing. For instance, I now run an effective altruist student organization at Harvard because I realized that even though I’m a comparatively bad leader and don’t enjoy it very much, it’s still very high-impact if I work hard enough at it. This isn’t to say that I’m actually trying yet, but I’ve gotten closer.

Using this distinction between pretending and actually trying, I would summarize a lot of effective altruism as “pretending to actually try”. As a social group, effective altruists have successfully noticed the pretending/actually-trying distinction. But they seem to have stopped there, assuming that knowing the difference between fake trying and actually trying translates into ability to actually try. Empirically, it most certainly doesn’t. A lot of effective altruists still end up satisficing—finding actions that are on their face acceptable under core EA standards and then picking those which seem appealing because of other essentially random factors. This is more likely to converge on good actions than what society does by default, because the principles are better than society’s default principles. Nevertheless, it fails to make much progress over what is directly obvious from the core EA principles. As a result, although “doing effective altruism” feels like truth-seeking, it often ends up being just a more credible way to pretend to try.

Below I introduce various ways in which effective altruists have failed to go beyond the social-satisficing algorithm of establishing some credibly acceptable alternatives and then picking among them based on essentially random preferences. I exhibit other areas where the norms of effective altruism fail to guard against motivated cognition. Both of these phenomena add what I call “epistemic inertia” to the effective-altruist consensus: effective altruists become more subject to pressures on their beliefs other than those from a truth-seeking process, meaning that the EA consensus becomes less able to update on new evidence or arguments and preventing the movement from moving forward. I argue that this stems from effective altruists’ reluctance to think through issues of the form “being a successful social movement” rather than “correctly applying utilitarianism individually”. This could potentially be solved by introducing an additional principle of effective altruism—e.g. “group self-awareness”—but it may be too late to add new things to effective altruism’s DNA.

Philosophical difficulties

There is [currently wide disagreement among effective altruists](#) on the correct framework for population ethics. This is crucially important for determining the best way to improve the world: different population ethics can lead to drastically different choices (or at least so we would expect *a priori*), and if the EA movement can’t converge on at least their instrumental goals, it will quickly fragment and lose its power. Yet there has been little progress towards discovering the correct population ethics (or, from a moral anti-realist standpoint, constructing arguments that will lead

to convergence on a particular population ethics), or even determining which ethics lead to which interventions being better.

Poor cause choices

Many effective altruists donate to [GiveWell's](#) top charities. All three of these charities work in global health. Is that because GiveWell knows that global health is the highest-leverage cause? No. It's because it was the only one with enough data to say anything very useful about. There's little reason to suppose that this correlates with being particularly high-leverage—on the contrary, heuristic but less rigorous arguments for causes like [existential risk prevention](#), [vegetarian advocacy](#) and [open borders](#) suggest that these could be even more efficient.

Furthermore, the our current “best known intervention” is likely to change (in a more cost-effective direction) in the future. There are two competing effects here: we might discover better interventions to donate to than the ones we currently think are best, but we also might run out of opportunities for the current best known intervention, and have to switch to the second. So far we seem to be in a regime where the first effect dominates, and there's no evidence that we'll reach a tipping point very soon, especially given how new the field of effective charity research is.

Given these considerations, it's quite surprising that effective altruists are donating to global health causes now. Even for those looking to use their donations to set an example, a donor-advised fund would have many of the benefits and none of the downsides. And anyway, donating when you believe it's not (except for example-setting) the best possible course of action, in order to make a point about figuring out the best possible course of action and then doing that thing, seems perverse.

Non-obviousness

Effective altruists often express surprise that the idea of effective altruism only came about so recently. For instance, my student group recently hosted Elie Hassenfeld for a talk in which he made remarks to that effect, and I've heard other people working for EA organizations express the same sentiment. But no one seems to be actually worried about this—just smug that they've figured out something that no one else had.

The “market” for ideas is at least somewhat efficient: most simple, obvious and correct things get thought of fairly quickly after it's possible to think them. If a meme as simple as effective altruism hasn't taken root yet, we should at least try to understand why before throwing our weight behind it. The absence of such attempts—in other words, the fact that non-obviousness doesn't make effective altruists worried that they're missing something—is a strong indicator against the “effective altruists are actually trying” hypothesis.

Efficient markets for giving

It's often claimed that “nonprofits are not a market for doing good; they're a market for warm fuzzies”. This is used as justification for why it's possible to do immense amounts of good by donating. However, while it's certainly true that most donors aren't explicitly trying to purchase utility, there's still a lot of money that is.

The [Gates Foundation](#) is an example of such an organization. They're effectiveness-minded and with \$60 billion behind them. 80,000 Hours has already [noted](#) that they've probably saved over 6 million lives with their vaccine programs alone—given that they've spent a relatively small part of their endowment, they must be getting a much better exchange rate than our current best guesses.

So why not just donate to the Gates Foundation? Effective altruists need a better account of the “market inefficiencies” that they're exploiting that Gates isn't. Why didn't the Gates Foundation fund the [Against Malaria Foundation](#), GiveWell's top charity, when it's in one of their main research areas? It seems implausible that the answer is simple incompetence or the like.

A general rule of markets is that if you don't know what your edge is, you're the sucker. Many effective altruists, when asked what their edge is, give some answer along the lines of “actually being strategic/thinking about utility/caring about results”, and stop thinking there. This isn't a compelling case: as mentioned before, it's not clear why no one else is doing these things.

Inconsistent attitude towards rigor

Effective altruists insist on extraordinary rigor in their charity recommendations—cf. for instance GiveWell's work. Yet for many ancillary problems—donating now vs. later, choosing a career, and deciding how “meta” to go (between direct work, earning to give, doing advocacy, and donating to advocacy), to name a few—they seem happy to choose between the not-obviously-wrong alternatives based on intuition and gut feelings.

Poor psychological understanding

John Sturm suggests, and I agree, that many of these issues are psychological in nature:

I think a lot of these problems take root a commitment level issue:

I, for instance, am thrilled about changing my mentality towards charity, not my mentality towards having kids. My first guess is that - from an EA and overall ethical perspective - it would be a big mistake for me to have kids (even after taking into account the normal EA excuses about doing things for myself). At least right now, though, I just don't care that I'm ignoring my ethics and EA; I want to have kids and that's that.

This is a case in which I'm not “being lazy” so much as just not trying at all. But when someone asks me about it, it's easier for me to give some EA excuse (like that having kids will make me happier and more productive) that I don't think is true - and then I look like I'm being a lazy or careless altruist rather than not being one at all.

The model I'm building is this: there are many different areas in life where I could apply EA. In some of them, I'm wholeheartedly willing. In some of them, I'm not willing at all. Then there are two kinds of areas where it looks like I'm being a lazy EA: those where I'm willing and want to be a better EA... and those where I'm not willing but I'm just pretending (to myself or others or both).

The point of this: when we ask someone to be a less lazy EA, we are (1) helping them do a better job at something they want to do, and (2) trying to make them either do more than they want to or admit they are “bad”.

In general, most effective altruists respond to deep conflicts between effective altruism and other goals in one of the following ways:

1. Unconsciously resolve the cognitive dissonance with motivated reasoning: “it’s clearly my comparative advantage to spread effective altruism through poetry!”
2. Deliberately and knowingly use motivated reasoning: “dear Facebook group, what are the best utilitarian arguments in favor of becoming an EA poet?”
3. Take the easiest “honest” way out: “I wouldn’t be psychologically able to do effective altruism if it forced me to go into finance instead of writing poetry, so I’ll become an effective altruist poet instead”.

The third is debatably defensible—though, for a community that purports to put stock in rationality and self-improvement, effective altruists have shown surprisingly little interest in self-modification to have more altruistic intentions. This seems obviously worthy of further work.

Furthermore, EA norms do not proscribe even the first two, leading to a group norm that doesn’t cause people to notice when they’re engaging in a certain amount of motivated cognition. This is quite toxic to the movement’s ability to converge on the truth. (As before, effective altruists are still better than the general population at this; the core EA principles are strong enough to make people notice the most obvious motivated cognition that obviously runs afoul of them. But that’s not nearly good enough.)

Historical analogues

With the partial exception of GiveWell’s [history of philanthropy project](#), there’s been no research into good historical outside views. Although there are no direct precursors of effective altruism (worrying in its own right; see above), there is one notably similar movement: communism, where the idea of “from each according to his ability, to each according to his needs” originated. Communism is also notable for its various abject failures. Effective altruists need to be more worried about how they will avoid failures of a similar class—and in general they need to be more aware of the pitfalls, as well as the benefits, of being an increasingly large social movement.

Aaron Tucker elaborates better than I could:

In particular, Communism/Socialism was a movement that was started by philosophers, then continued by technocrats, where they thought reason and planning could make the world much better, and that if they coordinated to take action to fix everything, they could eliminate poverty, disease, etc.

Marx totally got the “actually trying vs. pretending to try” distinction AFAICT (“Philosophers have only explained the world, but the real problem is to change it” is a quote of his), and he really strongly rails against people who unreflectively try to fix things in ways that make sense to the culture they’re starting from—the problem isn’t that the bourgeoisie aren’t trying to help people, it’s that the only conception of help that the bourgeoisie have is one that’s mostly epiphenomenal to actually improving the lives of the proletariat—giving them nice bourgeoisie

things like education and voting rights, but not doing anything to improve the material condition of their life, or fix the problems of why they don't have those in the first place, and don't just make them themselves.

So if Marx got the pretend/actually try distinction, and his followers took over countries, and they had a ton of awesome technocrats, it seems like it's the perfect EA thing, and it totally didn't work.

Monoculture

Effective altruists are not very diverse. The vast majority are white, "upper-middle-class", intellectually and philosophically inclined, from a developed country, etc. (and I think it skews significantly male as well, though I'm less sure of this). And as much as the multiple-perspectives argument for diversity is hackneyed by this point, it seems quite germane, especially when considering e.g. global health interventions, whose beneficiaries are culturally very foreign to us.

Effective altruists are not very humanistically aware either. EA came out of analytic philosophy and spread from there to math and computer science. As such, they are too hasty to dismiss many arguments as moral-relativist postmodernist fluff, e.g. that effective altruists are promoting cultural imperialism by forcing a Westernized conception of "the good" onto people they're trying to help. Even if EAs are quite confident that the utilitarian/reductionist/rationalist worldview is correct, the outside view is that really engaging with a greater diversity of opinions is very helpful.

Community problems

The discourse around effective altruism in e.g. the [Facebook group](#) used to be of fairly high quality. But as the movement grows, the traditional venues of discussion are getting inundated with new people who haven't absorbed the norms of discussion or standards of proof yet. If this is not rectified quickly, the EA community will cease to be useful at all: there will be no venue in which a group truth-seeking process can operate. Yet nobody seems to be aware of the magnitude of this problem. There have been some half-hearted attempts to fix it, but nothing much has come of them.

Movement building issues

The whole point of having an effective altruism "movement" is that it'll be bigger than the sum of its parts. Being organized as a movement should turn effective altruism into the kind of large, semi-monolithic actor that can actually get big stuff done, not just make marginal contributions.

But in practice, large movements and truth-seeking hardly ever go together. As movements grow, they get more "epistemic inertia": it becomes much harder for them to update on evidence. This is because they have to rely on social methods to propagate their memes rather than truth-seeking behavior. But people who have been drawn to EA by social pressure rather than truth-seeking take much longer to change their beliefs, so once the movement reaches a critical mass of them, it will become difficult for it to update on new evidence. As described above, this is already happening to effective altruism with the ever-less-useful Facebook group.

Conclusion

I've presented several areas in which the effective altruism movement fails to converge on truth through a combination of the following effects:

1. Effective altruists "stop thinking" too early and satisfice for "doesn't obviously conflict with EA principles" rather than optimizing for "increases utility". (For instance, they choose donations poorly due to this effect.)
2. Effective altruism puts strong demands on its practitioners, and EA group norms do not appropriately guard against motivated cognition to avoid them. (For example, this often causes people to choose bad careers.)
3. Effective altruists don't notice important areas to look into, specifically issues related to "being a successful movement" rather than "correctly implementing utilitarianism". (For instance, they ignore issues around group epistemology, historical precedents for the movement, movement diversity, etc.)

These problems are worrying on their own, but the lack of awareness of them is the real problem. The monoculture is worrying, but the lackadaisical attitude towards it is worse. The lack of rigor is unfortunate, but the fact that people haven't noticed it is the real problem.

Either effective altruists don't yet realize that they're subject to the failure modes of any large movement, or they don't feel motivation to do the boring legwork of e.g. engaging with viewpoints that your inside view says are annoying but that the outside view says are useful on expectation. Either way, this bespeaks worrying things about the movement's staying power.

More importantly, it also indicates an epistemic failure on the part of effective altruists. The fact that no one else within EA has done a substantial critique yet is a huge red flag. If effective altruists aren't aware of strong critiques of the EA movement, why aren't they looking for them? This suggests that, contrary to the emphasis on rationality within the movement, many effective altruists' beliefs are based on social, rather than truth-seeking, behavior.

If it doesn't solve these problems, effective-altruism-the-movement won't help me achieve any more good than I could individually. All it will do is add epistemic inertia, as it takes more effort to shift the EA consensus than to update my individual beliefs.

Are these problems solvable?

It seems to me that the third issue above (lack of self-awareness as a social movement) subsumes the other two: if effective altruism as a movement were sufficiently introspective, it could probably notice and solve the other two problems, as well as future ones that will undoubtedly crop up.

Hence, I propose an additional principle of effective altruism. In addition to being *altruistic*, *maximizing*, *egalitarian*, and *consequentialist* we should be *self-aware*: we should think carefully about the issues associated with being a successful movement, in order to make sure that we can move beyond the obvious applications of EA principles and come up with non-trivially better ways to improve the world.

Acknowledgments

Thanks to [Nick Bostrom](#) for coining the idea of a hypothetical apostasy, and to [Will Eden](#) for [mentioning](#) it recently.

Thanks to Michael Vassar, [Aaron Tucker](#) and Andrew Rettke for inspiring various of these points.

Thanks to Aaron Tucker and John Sturm for reading an advance draft of this post and giving valuable feedback.

Cross-posted from <http://www.benkuhn.net/ea-critique> since I want outside perspectives, and also LW's comments are nicer than mine.

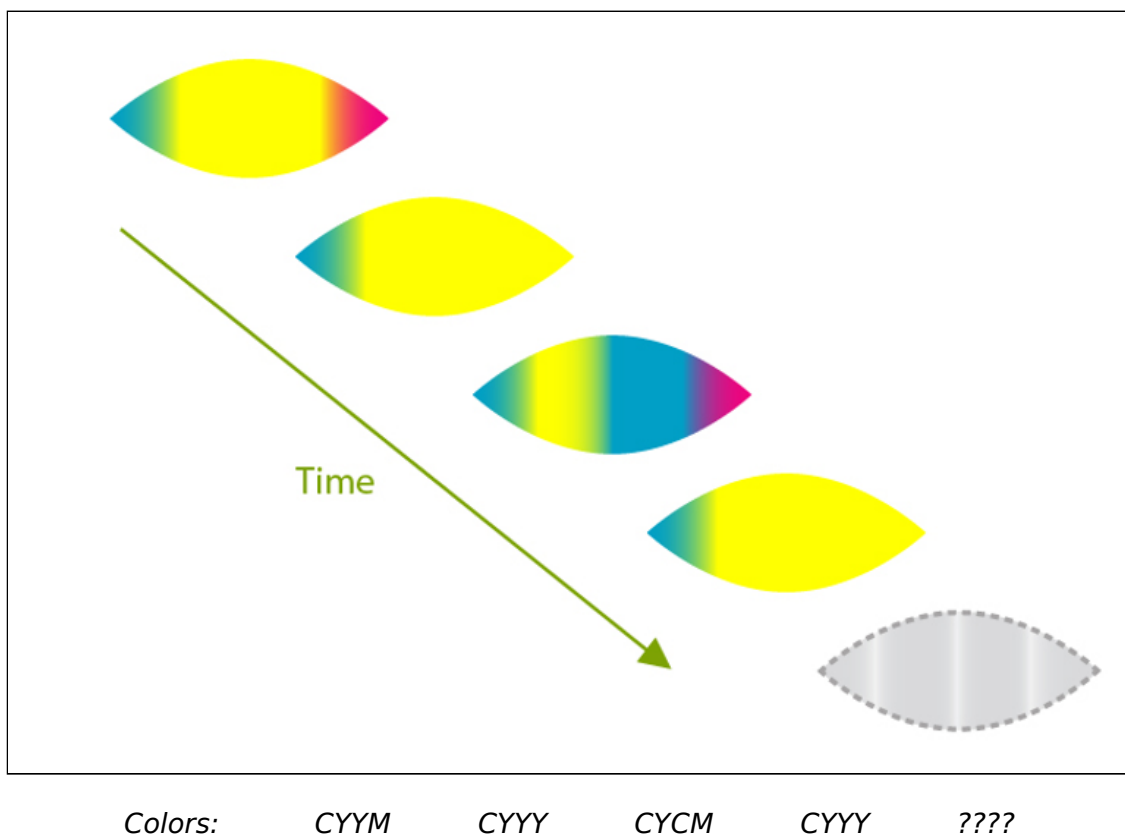
Building Phenomenological Bridges

[Naturalized induction](#) is an open problem in [Friendly Artificial Intelligence](#) (OPFAI). The problem, in brief: Our current leading models of induction do not allow reasoners to treat their own computations as processes in the world.

The problem's roots lie in algorithmic information theory and formal epistemology, but finding answers will require us to wade into debates on everything from theoretical physics to anthropic reasoning and self-reference. This post will lay the groundwork for a sequence of posts (titled '**Artificial Naturalism**') introducing different aspects of this OPFAI.

AI perception and belief: A toy model

A more concrete problem: Construct an algorithm that, given a sequence of the colors cyan, magenta, and yellow, predicts the next colored field.



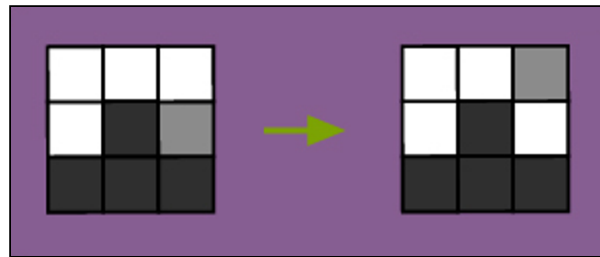
This is an instance of the general problem 'From an incomplete data series, how can a reasoner best make predictions about future data?'. In practice, any agent that acquires

information from its environment and makes predictions about what's coming next will need to have two map-like¹ subprocesses:

1. Something that generates the agent's predictions, its expectations. By analogy with human scientists, we can call this prediction-generator the agent's **hypotheses** or **beliefs**.
2. Something that transmits new information to the agent's prediction-generator so that its hypotheses can be [updated](#). Employing another anthropomorphic analogy, we can call this process the agent's **data** or **perceptions**.

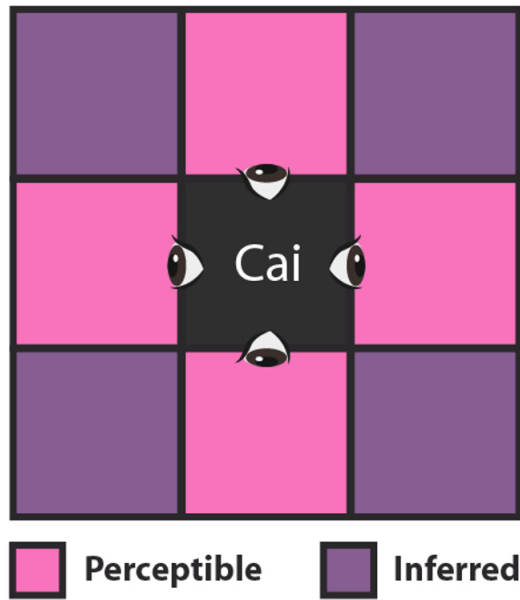
Here's an example of a hypothesis an agent could use to try to predict the next color field. I'll call the imaginary agent '**Cai**'. Any reasoner will need to begin with some (perhaps provisional) assumptions about the world.² Cai begins with the belief³ that its environment behaves like a [cellular automaton](#): the world is a grid whose tiles change over time based on a set of stable laws. The laws are local in time and space, meaning that you can perfectly predict a tile's state based on the states of the tiles next to it a moment prior — if you know which laws are in force.

Cai believes that it lives in a closed 3x3 grid where tiles have no diagonal effects. Each tile can occupy one of three states. We might call the states '0', '1', and '2', or, to make visualization easier, 'white', 'black', and 'gray'. So, on Cai's view, the world as it changes looks something like this:



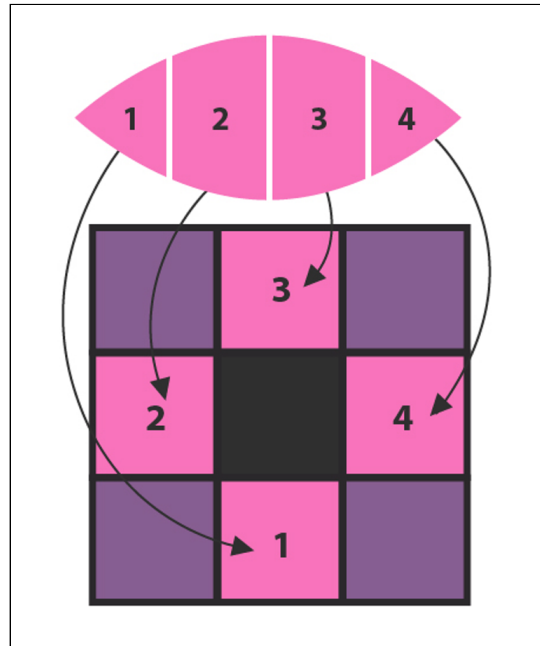
An example of the world's state at one moment, and its state a moment later.

Cai also has beliefs about its own location in the cellular automaton. Cai believes that it is a black tile at the center of the grid. Since there are no diagonal laws of physics in this world, Cai can only directly interact with the four tiles directly above, below, to the left, and to the right. As such, any perceptual data Cai acquires will need to come from those four tiles; anything else about Cai's universe will be known [only by inference](#).



Cai perceives stimuli in four directions. Unobservable tiles fall outside the cross.

How does all this bear on the color-predicting problem? Cai hypothesizes that the sequence of colors is sensory — it's an experience within Cai, triggered by environmental changes. Cai conjectures that since its visual field comes in at most four colors, its visual field's quadrants probably represent its four adjacent tiles. The leftmost color comes from a southern stimulus, the next one to the right from a western stimulus, then a northern one, then an eastern one. And the south, west, north, east cycle repeats again and again.

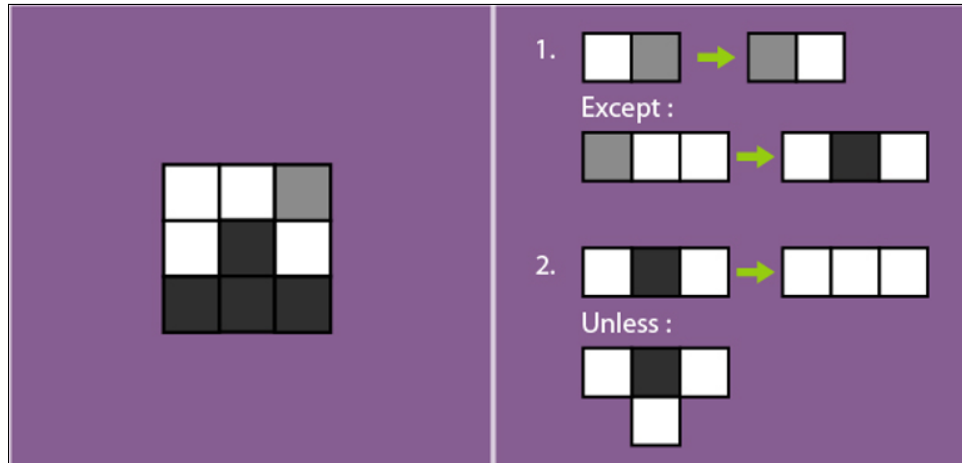


Cai's visual experiences break down into quadrants, corresponding to four directions.

On this model, the way Cai's senses organize the data isn't wholly veridical; the four patches of color aren't perfectly shaped like Cai's environment. But the organization of Cai's sensory apparatus and the organization of the world around Cai are similar enough that Cai can reconstruct many features of its world.

By linking its visual patterns to patterns of changing tiles, Cai can hypothesize laws that guide the world's changes and explain Cai's sensory experiences. Here's one possibility, Hypothesis A:

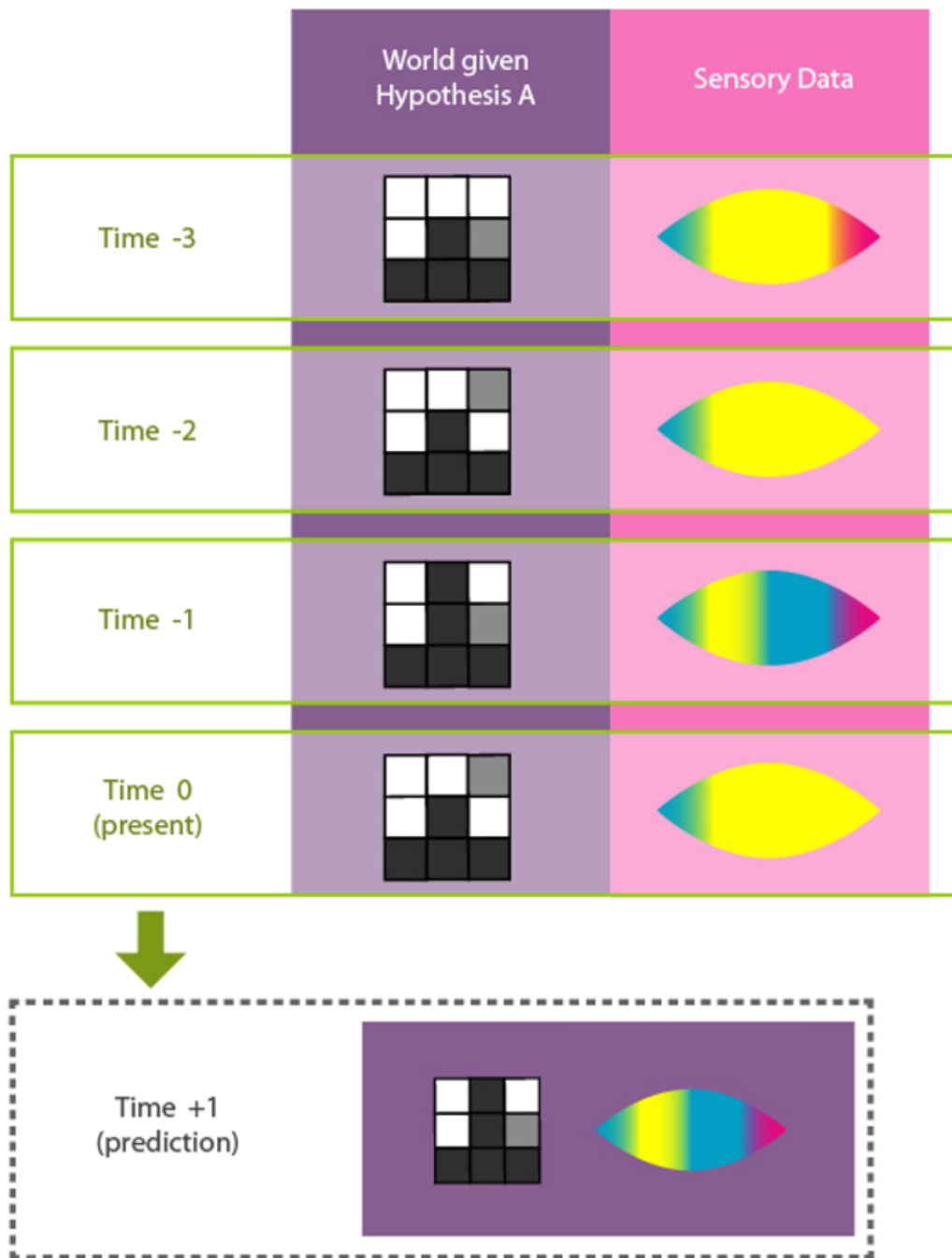
- Black corresponds to cyan, white to yellow, and gray to magenta.
- At present, the top two rows are white and the bottom row is black, except for the upper-right tile (which is gray) and Cai itself, a black middle tile.
- Adjacent gray and white tiles exchange shades. Exception: When a white tile is pinned by a white and gray tile on either side, it turns black.
- Black tiles pinned by white ones on either side turn white. Exception: When the black tile is adjacent to a third white tile, it remains black.



Hypothesis A's physical content. On the left: Cai's belief about the world's present state. On the right: Cai's belief about the rules by which the world changes over time. The rules are symmetric under rotation and reflection.

Bridging stimulus and experience

So that's one way of modeling Cai's world; and it will yield a prediction about the cellular automaton's next state, and therefore about Cai's next visual experience. It will also yield retrodictions of the cellular automaton's state during Cai's three past sensory experiences.











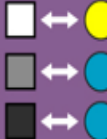




Hypothesis A asserts that tiles below Cai, to Cai's left, above, and to Cai's right relate to Cai's color experiences via the rule {black ↔ cyan, white ↔ yellow, gray ↔ magenta}. Corner tiles, and future world-states and experiences, can be inferred from Hypothesis A's cell transition rules.

Are there other, similar hypotheses that can explain the same data? Here's one, Hypothesis B:

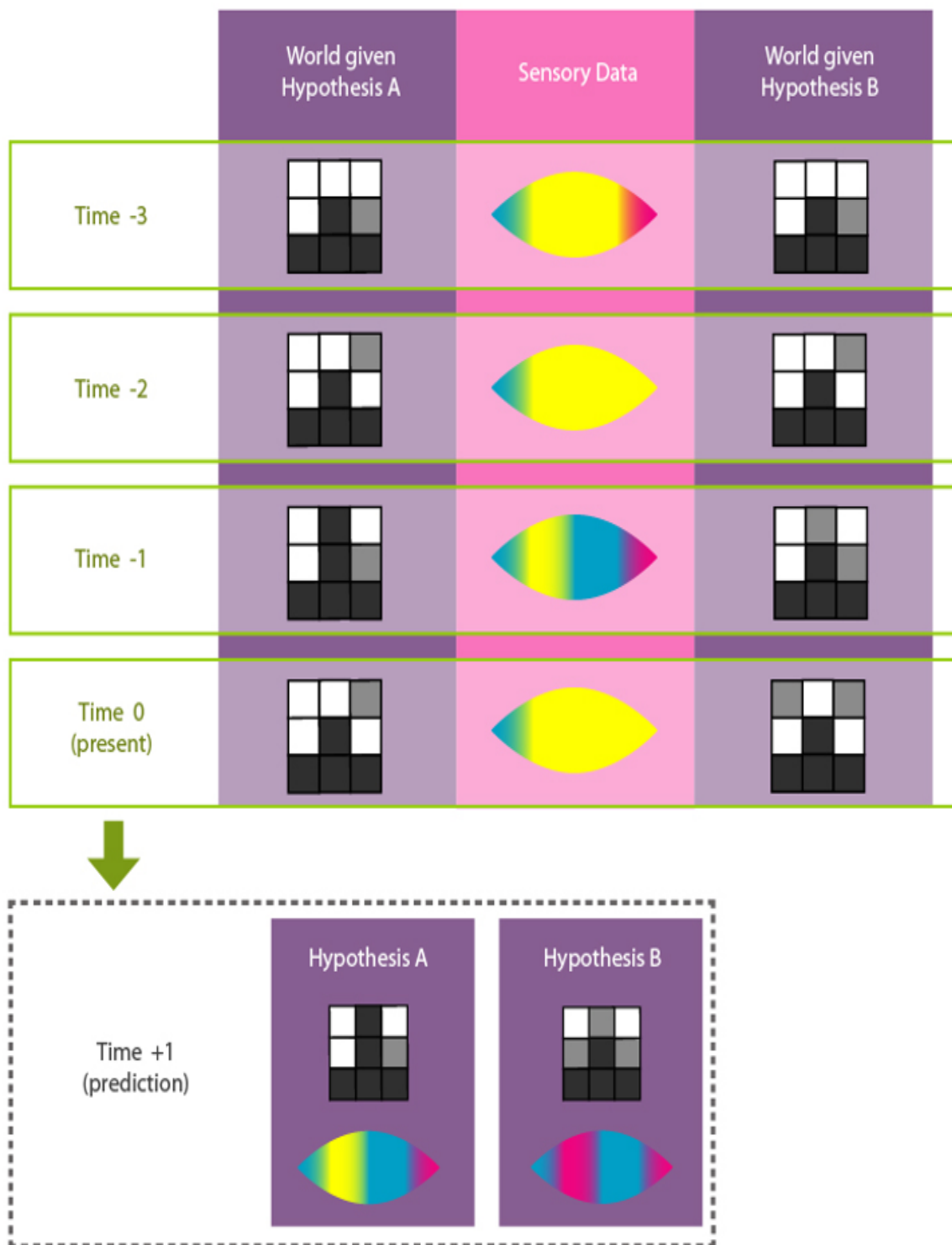
- Normally, the correspondences between experienced colors and neighboring tile states are {black ↔ cyan, white ↔ yellow, gray ↔ magenta}, as in Hypothesis A. But northern grays are perceived as though they were black, helping explain irregularities in the distribution of cyan.
- Hypothesis B's cellular automaton presently looks similar to Hypothesis A's, but with a gray tile in the upper-left corner.
- Adjacent gray and white tiles exchange shades. Nothing else changes.

The added complexity in the perception-to-environment link allows Hypothesis B to do away with most of the complexity in Hypothesis A's physical laws. Breaking down Hypotheses A and B into their respective physical and perception-to-environment components makes it more obvious how the two differ:

	Physical State	Physical Dynamic	Bridge Hypothesis
Hypothesis A		1.  Except :  2.  Unless : 	 
Hypothesis B		1. 	<div style="display: flex; justify-content: space-around;"> <div> Above :   </div> <div> Otherwise :   </div> </div>

A has the simpler bridge hypothesis, while B has the simpler physical hypothesis.

Though they share a lot in common, and both account for Cai's experiences to date, these two hypotheses diverge substantially in the cellular automaton states and future experiences they predict:



The two hypotheses infer different distributions and dynamical rules for the tile shades from the same perceptual data. These worldly differences then diverge in the future experiences they predict.

Hypotheses linking observations to theorized entities appear to be quite different from hypothesis that just describe the theorized entities in their own right. In Cai's case, the latter hypotheses look like pictures of physical worlds, while the former are ties between different kinds of representation. But in both cases it's useful to treat these processes in humans or machines as beliefs, since they can be assigned weights of expectation and updated.

'Phenomenology' is a general term for an agent's models of its own introspected experiences. As such, we can call these hypotheses linking experienced data to theorized processes **phenomenological bridge hypotheses**. Or just 'bridge hypotheses', for short.

If we want to build an agent that tries to evaluate the accuracy of a model based on the accuracy of its predictions, we need some scheme to compare thingies in the model (like tiles) and thingies in the sensory stream (like colors). Thus a **bridge rule** appears to be necessary to talk about induction over models of the world. And bridge hypotheses are just bridge rules treated as probabilistic, updatable beliefs.

As the last figure above illustrates, bridge hypotheses can make a big difference for one's scientific beliefs and [expectations](#). And bridge hypotheses aren't a free lunch; it would be a mistake to shunt all complexity onto them in order to simplify your physical hypotheses. Allow your bridge hypotheses to get too complicated, and you'll be able to justify mad world-models, e.g., ones where the universe consists of a single apricot whose individual atoms each get a separate bridge to some complex experience. At the same time, if you demand *too much* simplicity from your bridge hypotheses, you'll end up concluding that the physical world consists of a series of objects shaped just like your mental states. That way you can get away with a comically [simple](#) bridge rule like $\{ \text{exists}(x) \leftrightarrow \text{experiences}(y,x) \}$.

In the absence of further information, it may not be possible to rule out Hypothesis A or Hypothesis B. The takeaway is that tradeoffs between the complexity of bridging hypotheses and the complexity of physical hypotheses do occur, and do matter. Any artificial agent needs some way of formulating good hypotheses of this type in order to be able to understand the universe at all, whether or not it finds itself in doubt after it has done so.

Generalizing bridge rules and data

Reasoners — both human and artificial — don't begin with perfect knowledge of their own design. When they have working self-models at all, these self-models are fallible. Aristotle thought the brain was an organ for cooling the blood. We had to find out about neurons by opening up the heads of people who looked like us, putting the big corrugated gray organ under a microscope, seeing (with our eyes, our visual cortex, our senses) that the microscope (which we'd previously generalized shows us tiny things as if they were large) showed this incredibly fine mesh of connected blobs, and realizing, "Hey, I bet this does information processing and *that's* what I am! The big gray corrugated organ that's inside my own head is *me!*"

The bridge hypotheses in Hypotheses A and B are about linking an agent's environment-triggered experiences to environmental causes. But in fact bridge hypotheses are more general than that.

1. An agent's experiences needn't all have *environmental* causes. They can be caused by something inside the agent.
2. The cause-effect relation we're bridging can go the other way. E.g., a bridge hypothesis can link an experienced *decision* to a behavioral consequence, or to an expected outcome of the behavior.
3. The bridge hypothesis needn't link causes to effects at all. E.g., it can assert that the agent's experienced sensations or decisions just *are* a certain physical state. Or it can assert neutral correlations.

Phenomenological bridge hypotheses, then, can relate theoretical posits to any sort of experiential data. Experiential data are internally evident facts that get compared to hypotheses and cause updates — the kind of data of direct epistemic relevance to individual scientists updating their personal beliefs. Light shines on your retina, gets transduced to neural firings, gets reconstructed in your visual cortex and then — this is the key part — that internal fact gets used to decide what sort of universe you're probably in.

The data from an AI's environment is just one of many kinds of information it can use to update its probability distributions. In addition to ordinary sensory content such as vision and smell, update-triggering data could include things like how much RAM is being used. This is because an inner RAM sense can tell you that the universe is such as to include a copy of you with at least that much RAM.

We normally think of science as reliant mainly on sensory faculties, not [introspective](#) ones. Arriving at conclusions just by [examining your own intuitions and imaginings](#) sounds more like math or philosophy. But for present purposes the distinction isn't important. What matters is just whether the AGI forms accurate beliefs and makes good decisions. Prototypical scientists may shun [introspectionism](#) because humans do a better job of directly apprehending and communicating facts about their environments than facts about their own inner lives, but AGIs can have a very different set of strengths and weaknesses. Although introspection, like sensation, is fallible, introspective self-representations sometimes empirically [correlate](#) with world-states.⁴ And that's all it takes for them to constitute [Bayesian evidence](#).

Bridging hardware and experience

In my above discussion, all of Cai's world-models included representations of Cai itself. However, these representations were very simple — no more than a black tile in a specific environment. Since Cai's own computations are complex, it must be the case that either they are occurring outside the universe depicted (as though Cai is plugged into a cellular automaton [Matrix](#)), or the universe depicted is much more complex than Cai thinks.⁵ Perhaps its model is wildly mistaken, or perhaps the high-level cellular patterns it's hypothesized arise from other, smaller-scale regularities.

Regardless, Cai's computations must be embodied in some causal pattern. Cai will eventually need to construct bridge hypotheses between its experiences and their physical substrate if it is to make reliable predictions about its own behavior and about its relationship with its surroundings.

Visualize the epistemic problem that an agent needs to solve. Cai has access to a series of sensory impressions. In principle we could also add introspective data to that. But

you'll still get a series of (presumably time-indexed) facts in some native format of that mind. Those facts very likely won't be structured exactly like any ontologically basic feature of the universe in which the mind lives. They won't be a precise position of a Newtonian particle, for example. And even if we were dealing with sense data shaped just like ontologically basic facts, a rational agent could never [know for certain](#) that they were ontologically basic, so it would still have to consider hypotheses about even more basic particles.

When humans or AGIs try to match up hypotheses about universes to sensory experiences, there will be a type error. Our representation of the universe will be in hypothetical atoms or quantum fields, while our representation of sensory experiences will be in a native format like 'red-green'.⁶ This is where bridge rules like Cai's color conversions come in — bridges that relate our experiences to environmental stimuli, as well as ones that relate our experiences to the hardware that runs us.



Cai can form physical hypotheses about its own internal state, in addition to ones about its environment. This means it can form bridge hypotheses between its experiences and its own hardware, in addition to ones between its experiences and environment.

If you were an AI, you might be able to decode your red-green visual field into binary data — on-vs.-off — and make very simple hypotheses about how that corresponded to transistors making you up. Once you used a microscope on yourself to see the transistors, you'd see that they had binary states of positive and negative voltage, and all that would be left would be a hypothesis about whether the positive (or negative) voltage corresponded to an introspected 1 (or 0).

But even then, I don't quite see how you could do without the bridge rules — there has to be some way to go from internal sensory types to the types featured in your hypotheses about physical laws.

Our sensory experience of red, green, blue *is* certain neurons firing in the visual cortex, and these neurons are in turn made from atoms. But internally, so far as information processing goes, we just know about the red, the green, the blue. This is what you'd expect an agent made of atoms to feel like [from the inside](#). Our native representation of a pixel field won't come with a little tag telling us with infallible transparency about the underlying quantum mechanics.

But this means that when we're done positing a physical universe in all its detail, we also need one last (hopefully simple!) step that connects hypotheses about 'a brain that processes visual information' to 'I see blue'.

One way to avoid worrying about bridge hypotheses would be to instead code the AI to accept **bridge axioms**, bridge rules with no degrees of freedom and no uncertainty. But the AI's designers are not in fact [infinitely confident](#) about how the AI's perceptual states emerge from the physical world — that, say, quantum field theory is the One True Answer, and shall be so from now until the end of time. Nor can they transmit infinite rational confidence to the AI merely by making it more stubbornly convinced of the view. If you pretend to know more than you do, [the world will still bite back](#). As an agent in the world, you really do have to think about and test a variety of different uncertain hypotheses about what hardware you're running on, what kinds of environmental triggers produce such-and-such experiences, and so on. This is particularly true if your hardware is likely to undergo substantial changes over time.

If you don't allow the AI to form probabilistic, updatable hypotheses about the relation between its phenomenology and the physical world, the AI will either be unable to reason at all, or it will reason its way off a cliff. In my next post, [Bridge Collapse](#), I'll begin discussing how the latter problem sinks an otherwise extremely promising approach to formalizing ideal AGI reasoning: Solomonoff induction.

¹ By 'map-like', I mean that the processes look similar to the representational processes in human thought. They systematically correlate with external events, within a pattern-tracking system that can readily propagate and exploit the correlation. [↵](#)

² Agents need [initial assumptions](#), built-in [prior information](#). The prior is defined by whatever algorithm the reasoner follows in making its very first updates.

If I leave an agent's priors undefined, no [ghost](#) of [reasonableness](#) will intervene to give the agent a '[default](#)' prior. For example, it won't default to a uniform prior over possible coinflip outcomes in the absence of relevant evidence. Rather, without something that acts like a prior, the agent just won't work — in the same way that a calculator won't work if you grant it the freedom to do math however it wishes. A [frequentist](#) AI might refuse to *talk* about priors, but it would still need to *act* like it has priors, else break. [↵](#)

³ This talk of 'belief' and 'assumption' and 'perception' *is* anthropomorphizing, and the analogies to human psychology won't be perfect. This is important to keep in view, though there's only so much we can do to avoid vagueness and analogical reasoning when the architecture of AGIs remains unknown. In particular, I'm not assuming that every artificial scientist is particularly intelligent. Or particularly conscious.

What I mean with all this 'Cai believes...' talk is that Cai weights predictions and selects actions *just as though* it believed itself to be in a cellular automaton world. One can treat Cai's automaton-theoretic model as just a bookkeeping device for assigning [Cox's-theorem-following](#) real numbers to encoded images of color fields. But one can also treat Cai's model as a psychological expectation, to the extent it functionally resembles the corresponding human mental states. Words like 'assumption' and 'thinks' here needn't mean that the agent thinks in the same fashion humans think; what we're interested in are the broad class of information-processing algorithms that yield similar [behaviors](#). ↵

⁴ To illustrate: In principle, even a human pining to become a parent could, by introspection alone, infer that they might be an evolved mind (since they are experiencing a desire to self-replicate) and embedded in a universe which had evolved minds with evolutionary histories. An AGI with more reliable internal monitors could learn a great deal about the rest of the universe just by investigating itself. ↵

⁵ In either case, we shouldn't be surprised to see Cai failing to fully represent its own inner workings. An agent cannot explicitly represent itself in its totality, since it would then need to represent itself representing itself representing itself ... ad infinitum. Environmental phenomena, too, must usually be [compressed](#). ↵

⁶ One response would be to place the blame on Cai's positing white, gray, and black for its world-models, rather than sticking with cyan, yellow, and magenta. But there will still be a type error when one tries to compare perceived cyan/yellow/magenta with hypothesized (but perceptually invisible) cyan/yellow/magenta. Explicitly introducing separate words for hypothesized v. perceived colors doesn't produce the distinction; it just makes it easier to keep track of a distinction that was already present. ↵

Review of Scott Adams' "How to Fail at Almost Everything and Still Win Big"

Dilbert creator and bestselling author Scott Adams recently wrote a LessWrong compatible advice book that even contains a long list of cognitive biases. Adams told me in a phone interview that he is a lifelong consumer of academic studies, which perhaps accounts for why his book jibes so well with LessWrong teachings. Along with [HPMOR](#), [How to Fail at Almost Everything and Still Win Big](#) should be among your first choices when recommending books to novice rationalists. Below are some of the main lessons from the book, followed by a summary of my conversation with Adams about issues of particular concern to LessWrong readers.

My favorite passage describes when Adams gave a talk to a fifth-grade class and asked everyone to finish the sentence "If you play a slot machine long enough, eventually you will..." The students all shouted "WIN!" because, Adams suspects, they had the value of persistence drilled into them and confused it with success.

"WIN!" would have been the right answer if you didn't have to pay to play but the machine still periodically gave out jackpots. Adams thinks you can develop a system to turn your life into a winning slot machine that doesn't require money but does require "time, focus, and energy" to repeatedly pull the lever.

Adams argues that maximizing your energy level through proper diet, exercise, and sleep should take priority over everything else. Even if your only goal is to help others, be selfish with respect your energy level because it will determine your capacity for doing good. Adams has convinced me that appropriate diet, exercise, and sleep should be the starting point for effective altruists. Adams believes we have limited willpower and argues that if you make being active every single day a habit, you won't have to consume any precious willpower to motivate yourself to exercise.

Since most pulls of the life slot machine will win you nothing, Adams argues that lack of fear of embarrassment is a key ingredient for success. Adams would undoubtedly approve of [CFAR's comfort zone expansion exercises](#).

Adams lists skills that increase your chances of success. These include knowledge of public speaking, psychology, business writing, accounting, design, conversation, overcoming shyness, second language, golf, proper grammar, persuasion, technology (hobby level), and proper voice technique. He gives a bit of actionable advice on each, basically ideas for [becoming more awesome](#). I wish my teenage self had been told of Adams' theory that a shy person can frequently achieve better social outcomes by pretending that he is an actor playing the part of an extrovert.

Adams believes we should rely on systems rather than goals, and indeed he thinks that "goals are for losers." If, when playing the slot machine of life, your goal is winning a jackpot, then you will feel like a loser each time you don't win. But if, instead, you are systems oriented then you can be in the constant state of success, something that will probably make you happier.

Adams claims that happiness “isn’t as dependent on your circumstances as you might think,” and “anyone who has experienced happiness probably has the capacity to spend more time at the top of his or her personal range and less time near the bottom.” His suggestions for becoming happier include improving your exercise, diet, and sleep; having a flexible work schedule, being able to imagine a better future, and being headed towards a brighter future.

The part of the book most likely to trouble LessWrong readers is when Adams recommends engaging in self-delusion. He writes:

“Athletes are known to stop shaving for the duration of a tournament or to wear socks they deem lucky. These superstitions probably help in some small way to bolster their confidence, which in turn can influence success. It’s irrelevant that lucky socks aren’t a real thing...Most times our misconceptions about reality are benign and sometimes, even helpful. Other times, not so much.”

For me, being rational means having accurate beliefs and useful emotions. But what if these two goals conflict? For example, a college friend of mine who used to be a book editor [wrote](#) “Most [authors] would do better by getting a minimum wage job and spending their entire paycheck on lottery tickets.” I know this is true for me, yet I motivated myself to write [my last book](#) in part by repeatedly dreaming of how it would be a best seller, and such willful delusions did, at the very least, make writing the book more enjoyable.

To successfully distort reality you probably need to keep two separate mental books: the false one designed to motivate yourself, and the accurate one to keep you out of trouble. If you forget that you don’t really have a “[reality distortion field](#)”, that you can’t change the territory by falsifying your map, you might make a [Steve Jobs level error](#) by, say, voluntarily forgoing lifesaving medical care because you think you can wish your problem away.

The strangest part of the book concerns affirmations, which Adams defines as the “practice of repeating to yourself what you want to achieve while imagining the outcome you want.” Adams describes his fantastic success with achieving his affirmations, which included his becoming a famous cartoonist, having a seemingly hopeless medical problem fixed, and scoring at exactly the 94th percentile on the GMATs. Adams writes that the success of affirmations for him and others seems to go beyond what could be achieved by positive thinking. Thankfully he rules out magic as a possible solution and suggests that the success of affirmations might be due to selective memories, false memories, optimists tending to notice opportunities, selection effect of people who make affirmations, and mysterious science we don’t understand. His support of affirmations seems to contradict his dislike of goals.

Our Phone Conversation

I took advantage of [this offer](#) to get a 15-minute phone interview with Adams.

He has heard of LessWrong, but doesn’t have any specific knowledge of us. He thinks the Singularity is a very probable outcome for mankind. He believes it will likely turn out all right due to what he calls “[Adams’ Law of Slow-Moving Disasters](#)” which says that “any disaster we see coming with plenty of advance notice gets fixed.” To the extent that Adams is correct, we will owe much to [Eliezer and associates](#) for providing us with loud early warnings of how the Singularity might go wrong.

I failed in my brief attempt to get Adams interested in cryonics. He likes the idea of brain uploading and thinks it will be unnecessary to save the biological part of us. I was unable to convince him that cryonics is a good intermediate step until someone develops the technology for uploading. He mentioned that so long as the Internet survives a huge amount of him basically will as well.

Recalling Adams' claim that having a high tolerance for embarrassment is a key attribute for success, I asked him about my theory of how American dating culture, in which it's usually the man who risks rejection by asking the woman to go on the first date, gives men an entrepreneurial advantage because it eventually makes men more tolerant of rejection. Adams didn't directly respond to my theory, but brought up evolutionary psychology by saying that men would encounter much rejection because of their preference for variety and role as hunters. Adams stressed that this was just speculation and not back up by evidence.

Adams has heard of [MetaMed](#). He is very optimistic about the ability of medicine to become more rational and to handle big data. When I pointed out that doctors often [don't know basic statistics](#), he said that this probably doesn't impede their ability to treat patients.

After I explained the concept of akrasis to Adams, he mentioned the book "The Power of Habit" and told me that we can use science to develop better habits. (Kaj Sotala recently wrote a highly upvoted LessWrong [post](#) on the "The Power of Habit.")

Adams suggested that if you have trouble accomplishing a certain task, just focus on the part that you can do: for example if spending an hour at the gym seems too difficult, then just think about putting on your sneakers.

Although he didn't use these exact words, Adams basically defended himself against the charge of "[other-optimizing](#)." He explained that it would be very difficult to describe to an alien what a horse is. But once you succeeded describing the horse, it would be much easier to describe a zebra because you could do so in part by making references to the horse. Adams said he knows his advice isn't ideal for everyone, but it provides a useful template you can use to develop a plan better optimized for yourself.

At the end of the interview Adams said he was surprised I had not brought up assisted suicide, given his recent blog post on the topic. In the [post](#) Adams wrote:

"My father, age 86, is on the final approach to the long dirt nap (to use his own phrase). His mind is 98% gone, and all he has left is hours or possibly months of hideous unpleasantness in a hospital bed. I'll spare you the details, but it's as close to a living Hell as you can get..."

"If you're a politician who has ever voted against doctor-assisted suicide, or you would vote against it in the future, I hate your fucking guts and I would like you to die a long, horrible death. I would be happy to kill you personally and watch you bleed out. I won't do that, because I fear the consequences. But I'd enjoy it, because you motherfuckers are responsible for torturing my father. Now it's personal."

Based on [this blog post](#) I suspect that Yvain would agree with Adams about the magnitude and intensity of the evil of outlawing assisted suicide for the terminally ill.

If you are unwilling to buy Adams' book, I strongly recommend you at least [read his blog](#), which normally has a much softer and less angry tone than the passage I just cited.

Meditation: a self-experiment

Introduction

The LW/CFAR community has a fair amount of interest in meditation. This isn't surprising; many of the people who practiced and wrote about meditation in the past were trying to train a skill similar to rationality. Schools of meditation seem to be the closest already-existing thing to rationality dojos—this doesn't mean that they're very similar, only that I can't think of anything else that's more similar.

People are Doing Science on meditation; there are studies on the effects of meditation on [attention](#), [depression](#), [anxiety](#), [stress and pain reduction](#). [Insert usual disclaimer that many of these studies either won't be replicated or aren't measuring what they think they're measuring]. Meditation is apparently considered a [form of alternative medicine](#); this is quite annoying, actually, since it's a thing that might help a lot of people being lumped in with other things that almost certainly don't work.

[There's the spiritual enlightenment element of meditation, too. I won't touch on that, since my own experience isn't related to that aspect.]

Brienne Strohl has posted about [meditation and metacognition](#); DavidM has posted on [meditation and insight](#). Valentine, of CFAR, talked about mindfulness meditation helping to dispel the illusion of being hurried and never having enough time.

In short, lots of hype—enough that I found it worthwhile to give it a try myself. The main benefit I hoped to attain from practicing meditation was better control of attention—to be able to aim my attention more reliably at a particular target, and notice more quickly when it drifted. The secondary benefit would be better understanding and control of emotions, which I had already tried to accomplish through techniques other than meditation. However, I'd had the experience for several years of thinking that meditation was a valuable thing to try, and not trying it—evidence that I needed more than good intentions.

The experiment

Sometime in early September, I saw a poster on the wall at the hospital where I work, advertising a study on mindfulness meditation for people with social anxiety. I called the number on the poster and got myself enrolled because it was a good pre-commitment strategy. The benefits were deadlines, social pressure, and structure, with a steady supply of exercises, audio recordings, and readings. This came at the cost of two hours a week for twelve weeks, not all of which was spent on the specific skills that I wanted to learn. Another possible cost could be thinking of myself more as someone who has social anxiety, which might become a self-fulfilling prophecy, but I don't think this actually happened. If anything, sitting down in a group once a week with people whose anxiety significantly affected their functioning had the effect of making my own anxiety seem pretty insignificant. (I was able to convincingly make the case that I suffer from social anxiety during my interview; I've cried in front of my teachers a lot, including during my last year of nursing school, which caused some adults to think that I wasn't cut out for nursing).

I didn't quite do *all* of the homework for the study, which would have amounted to almost an hour a day. The social pressure of having to hand in a sheet had me doing most of it, though. I Beeminded twenty minutes a day of meditation; according to Beeminder, this has amounted to about 25 hours since mid-September despite the several occasions on which I derailed.

The Dropbox folder with the audio files for most of the meditation exercises I've done regularly is [here](#).

Breakdown of different meditation exercises

Compassionate Body Scan: a 25-minute tape with a man talking soporifically about exploring your feet like you would the feet of a beloved child and being curious about the experience of your ankles—and, eventually, the rest of your body. I've often done this one in bed when I hadn't gotten around to meditating earlier that day, and I often fall asleep around the pelvis area. I wish I could do this on demand; without the tape, it generally takes me 45 minutes to an hour to fall asleep.

10 and 20 minute sitting meditation: A walkthrough of focusing on the breath in various places; the nostrils, the throat, the chest and abdomen; and later focusing on the whole body. I like this one because sometimes near the end I feel like I'm floating in a void. I also do it in a particular position—kneeling and supporting my bum on a low stool—which I've conditioned myself to associate with meditation to the point that the posture is calming in and of itself.

Loving-Kindness meditation: Walks through feeling kindness towards someone you love, someone you're indifferent to, and someone you dislike. I don't usually feel very different after this one, but this is partly because I've been training myself to like people in general for years, and because nursing school in and of itself is an exercise in empathy-building. I have noticed that I can't do it as effectively anymore because there's no one I experience a strong emotion of dislike towards—I used a particular nurse on my unit for a month or so, and now, although I have the same thoughts about her, I don't have the actual emotional experience of dislike. So I guess it worked.

Mountain Meditation: a complex visualization/metaphor of yourself as a mountain. I would say that your mileage may vary the most on this one, because of variations in mental imagery. I have very vivid mental imagery, so I like it a lot. I've got some salient mental images cached now to draw on the metaphor of "I am a mountain and I will endure seasons, storms, winter, every single alarm in my patient's room going off at once, and WHATEVER ELSE THE UNIVERSE WANTS TO THROW AT ME!"

The Results

Initial positive; I like meditation enough to want to keep doing it. It feels good, overall. That doesn't mean that I want to do it all the time, or that I effortlessly accomplish my Beeminder goal; it does require willpower to put away the computer/book/food/music and focus on doing nothing for 20 minutes, and I do moan and groan and put it off. But it's generally pleasant while I'm actually doing it, and there are times when I feel a *lot* better afterwards.

There are several reasons why, a priori, I would expect meditation to be especially helpful for me. My natural state is to daydream. I'm good at remembering complex

sequences of things, but not at noticing details, because I'm too busy thinking about all those complex interesting things that happened earlier. This is all very well, but as a nurse, it's important for me to be paying attention to what I'm doing.

The single most helpful time for me to meditate is when I'm feeling very frazzled. This usually happens when I've had an extremely busy day at work, running around all day trying to save someone's life, and I feel very motivated, but at some point I get overwhelmed by all the object-level tasks that keep flying at my face, and I lose track of the "big picture" and with it the ability to prioritize or plan anything, and end up dealing with tasks in the order of which thing beeps the loudest. Even once I get home after a day like this, the frazzled state persists and I can't actually settle onto any tasks that need to get done at home. A friend of mine once described his experience of having tried cocaine as "I felt very alert, but it was an illusion because I was also really scattered." I haven't tried cocaine, but those words describe the me-after-a-busy-shift very accurately.

I don't always get myself to meditate as soon as I get home (or on break during the busy shift); it takes willpower, and this is a willpower-reduced state. It would be an excellent habit to train, though. To clarify: I find meditation *really difficult* in this state. My thoughts are racing and the last thing I want to focus on is my breath, because I did exciting things today and I should think about all of them really fast. But at least meditation forces me to focus on the fact that my thoughts are racing, and notice that from a calm perspective, instead of completely identifying with and being caught up in the flow. Twenty minutes later, I'm generally reset and able to do something else, although that thing is most often sleep.

The biggest overall change I've noticed after I started meditating regularly is more awareness of my physical and emotional states; physical especially. It's easy for me not to get around to drinking any water at work, for example, and then ten hours later noticing that I "don't feel well" in some vague way, but experiencing this mostly as the phrase "I don't feel well" in my head, as opposed to focusing on any physical sensation that might clue me in to the source of my discomfort. (Of course, outside view puts drinking water on the list of things I should try if I mysteriously don't feel well, but it's nice to have an actual physical sensation, too). Several of the meditation exercises had aspects of focusing on the body, focusing on sensations of discomfort without trying to apply words to them or interpret them, etc. This is a [5-second-level](#) skill that I've improved hugely at.

A specific instance is when I suddenly clue in that something very urgent and serious is happening, and I go from my general at-work state of mild anxiety to a full-blown SNS fight-or-flight adrenaline response. If I pay attention to my thoughts, they generally aren't going anywhere useful. But meanwhile, my body is doing all these interesting things; racing heart, shaky hands, that weird sinking crampy feeling in my stomach, etc. Meditation has trained me to automatically notice and pay attention to the physical sensations, which gives me a couple of seconds to get my thoughts calmed down. I also have a much easier time getting rid of the annoying physical effects like shaking hands, which make it hard to do fiddly things like draw medications up into syringes—and heaven forbid I ever have to try to put an IV in on a patient in cardiac arrest.

I identify less with my moods. I was already generally good at recognizing that moods were temporary coloured glasses on the world and not just the way the world actually was, but I'm better now. I can sometimes notice a negative mood and also notice the thing I actually have to do to fix it; this might be as simple as eating something, or

might involve observing my thoughts and realizing that the bad mood started unnoticed because of an interaction with someone else which I interpreted negatively. This is a skill that was discussed a great deal in my meditation class, and it's actually not where I've improved the most; the biggest change has been in the level of physical awareness.

The skills of body scanning and focusing on breathing have been helpful for forms of exercise that I find aversive, such as running. I can notice the cramps in my shins and explore them, rather than getting caught up in the mental verbal loop of "I have cramps in my shins and this is awful and I want it to stop!" Surprisingly, this helps. By focusing on my breathing in a meditation-y way while running; by this, I mean literally focusing on the cold feeling of the air going into my nostrils and throat and the feel of my clothes stretching over my stomach as it expands; I somehow clued in that my diaphragm actually could move independently of my legs and I could breathe at a normal rhythm and depth.

One of the homework exercises was executing certain activities "mindfully". I quickly learned which things I could do mindfully without changing my schedule much, and I noticed that various things just felt better. Swimming, for example, is *really* sensual when you are actually paying attention to the sensation of water flowing over your skin and the sound of it in your ears through a swim cap and the patterns of reflected light through foggy goggles. I have memories of my 11-year-old self experiencing this; at some point, I stopped. I spend a lot of my life on automatic. This isn't always a problem, but when I notice a negative mood, I can turn on more sensory experience and often get rid of it that way.

I've gotten somewhat better at actually experiencing myself as a modular mind, with various voices that want different things for different reasons. I don't have to endorse or identify with or experience myself as one of the voices; they're all me, and none of them are the "real" me. This helps with mental clarity, and being able to think about difficult decisions without actually experiencing the agonizing and aversiveness and confusion; yes, the different parts of me disagree with each other, that's just a fact about the state of the world and it's okay and doesn't mean I have to be angry.

Conclusion

The science on it predicts that meditation has positive expected value as a thing to try. My personal experience showed it to be an overall positive for me in particular; not life-changing, but worth spending 20 minutes a day on. Your mileage may vary, and anecdotally there's a lot of variation in how much people value get from it, but it doesn't take a lot of effort to try. I pre-committed to a twelve-week experiment, but it's likely possible to get an idea of whether meditation helps or not in a much shorter time span.

Speaking from my personal experience, meditation is likely to be helpful if you tend to daydream and are in a position where you need to daydream less. Science also says that meditation will probably help if you suffer a lot of distress from rumination or mulling over unpleasant thoughts.

If I were to repeat the experiment, I would do it with actual mood tracking, so that the data I got was more accurate than "In hindsight and upon reflection, I think I feel this way." I still haven't found a good method of mood tracking that does all the things I want it to.

Meditation is a clearly defined activity; there are groups of people who do it together, books about how to do it, and guided-meditation recordings that you can download from the Internet. If having structure helps you to actually do things, there are plenty of ways to obtain structure.

I am happy to discuss meditation further with anyone who is interested.

How the Grinch Ought to Have Stolen Christmas

On Dec. 24, 1957, a Mr. T. Grinch attempted to disrupt Christmas by stealing associated gifts and decorations. His plan failed, the occupants of Dr. Suess' narrative remained festive, and Mr. Grinch himself succumbed to cardiac hypertrophy. To help others avoid repeating his mistakes, I've written a brief guide to properly disrupting holidays. Holiday-positive readers should read this with the orthogonality thesis in mind. Fighting Christmas is tricky, because the obvious strategy - making a big demoralizing catastrophe - doesn't work. No matter what happens, the media will put the word *Christmas* in front of it and convert your scheme into even more free advertising for the holiday. It'll be a *Christmas* tragedy, a *Christmas* earthquake, a *Christmas* wave of foreclosures. That's no good; attacking Christmas takes more finesse.

The first thing to remember is that, whether you're stealing a holiday or a magical artifact of immense power, it's almost always a good idea to leave a decoy in its place. When people notice that something important is missing, they'll go looking to find or replace it. This rule can be generalized from physical objects to abstractions like *sense of community*. T. Grinch tried to prevent community gatherings by vandalizing the spaces where they would've taken place. A better strategy would've been to promise to organize a Christmas party, then skip the actual organizing and leave people to sit at home by themselves. Unfortunately, this solution is not scalable, but someone came up with a very clever solution: encourage people to watch Christmas-themed films instead of talking to each other, achieving almost as much erosion of community without the backlash.

I'd like to particularly applaud Raymond Arnold, for inventing a vaguely-Christmas-like holiday in December, with no gifts, and death (rather than cheer) as its central theme [1]. I really wish it didn't involve so much singing and community, though. I recommend raising the musical standards; people who can't sing at studio-recording quality should not be allowed to sing at all.

Gift-giving traditions are particularly important to stamp out, but stealing gifts is ineffective because they're usually cheap and replaceable. A better approach would've been to promote giving undesirable gifts, such as religious sculptures and fruitcake. Even better would be to convince the Mayor of Whoville to enact bad economic policies, and grind the Whos into a poverty that would make gift-giving difficult to sustain. Had Mr. Grinch pursued this strategy effectively, he could've stolen Christmas *and Birthdays* and gotten himself a Nobel Prize in Economics [2].

Finally, it's important to avoid rhyming. This is one of those things that should be completely obvious in hindsight, with a little bit of genre savvy; villains like us win much more often in prose and in life than we do in verse.

And with that, I'll leave you with a few closing thoughts. If you gave presents, your friends are disappointed with them. Any friends who didn't give you presents, it's because they don't care, and any fiends who did give you presents, they're cheap and lame presents for the same reason. If you have a Christmas tree, it's ugly, and if it's snowing, the universe is trying to freeze you to death.

Merry Christmas!

[1] I was initially concerned that the Solstice would pattern-match and mutate into a less materialistic version of Christmas, but running a Kickstarter campaign seems to have addressed that problem.

[2] This is approximately the reason why Alfred Nobel specifically opposed the existence of that prize.

Doubt, Science, and Magical Creatures - a Child's Perspective

Doubt

I grew up in a Jewish household, so I didn't have Santa Claus to doubt - but I did have the tooth fairy.

It was hard for me to believe that a magical being I had never seen somehow knew whenever any child lost their tooth, snuck into their house unobserved without setting off the alarms, for unknown reasons took the tooth, and for even less fathomable reasons left a dollar and a note in my mom's handwriting.

On the other hand, the alternative hypothesis was no less disturbing: my parents were lying to me.

Of course I had to know which of these terrible things was true. So one night, when my parents were out (though I was still young enough to have a babysitter), I noticed that my tooth was coming out and decided that this would be...

A Perfect Opportunity for an Experiment.

I reasoned that if my parents didn't know about the tooth, they wouldn't be able to fake a tooth fairy appearance. I would find a dollar and note under my pillow [if, but only if](#), the tooth fairy were real.

I solemnly told the babysitter, "I lost my tooth, but don't tell Mom and Dad. It's important - it's science!" Then at the end of the night I went to my bedroom, put the tooth under the pillow, and went to sleep. The next morning, I woke up and looked under my pillow. The tooth was gone, and in place there was a dollar and a note from the "tooth fairy."

This could have been the end of the story. I could have decided that I'd performed an experiment that would come out one way if the tooth fairy were real, and a different way if the tooth fairy were not. But I was more skeptical than that. I thought, "What's more likely? That a magical creature took my tooth? Or that the babysitter told my parents?"

I was furious at the possibility of such an egregious violation of experimental protocol, and never trusted that babysitter in the lab again.

An Improvement in Experimental Design

The next time, I was more careful. I understood that the flaw in the previous experiment had been failure to adequately conceal the information from my parents. So the next time I lost a tooth, I told no one. As soon as I felt it coming loose in my mouth, I ducked into the bathroom, ran it under the tap to clean it, wrapped it in a tissue, stuck it in my pocket, and went about my day as if nothing had happened. That night, when no one was around to see, I put the tooth under my pillow before I went to sleep.

In the morning, I looked under the pillow. No note. No dollar. Just that tooth. I grabbed the incriminating evidence and burst into my parents bedroom, demanding to know:

"If, as you say, there is a tooth fairy, then how do you explain THIS?!"

What can we learn from this?

The basic idea of the experiment was ideal. It was testing a binary hypothesis, and was expected to perfectly distinguish between the two possibilities. However, if I had known then what I know now about rationality, I could have done better.

As soon as my first experiment produced an unexpected positive result, just by learning that fact, I knew why it had happened, and what I needed to fix in the experiment to produce strong evidence. Prior to the first experiment would have been a perfect opportunity to apply the "Internal Simulator," as CFAR calls it - imagining in advance getting each of the two possible results, and what I think afterwards - do I think the experiment worked? Do I wish I'd done something differently? - in order to give myself the opportunity to correct those errors in advance instead of performing a costly experiment (I had a limited number of baby teeth!) to find them.

[Cross-posted](#) at my [personal blog](#).