Best of LessWrong: January 2013

1. 2012: Year in Review
2. Farewell Aaron Swartz (1986-2013)
3. Morality is Awesome
4. AidGrade - GiveWell finally has some competition
5. Assessing Kurzweil: the results
6. Course recommendations for Friendliness researchers
7. My simple hack for increased alertness and improved cognitive functioning: very bright light
8. The Zeroth Skillset
9. I attempted the AI Box Experiment (and lost)
10. Just One Sentence
11. Update on Kim Suozzi (cancer patient in want of cryonics)
12. Best of Rationality Quotes, 2012 Edition

Best of LessWrong: January 2013

1. 2012: Year in Review
2. Farewell Aaron Swartz (1986-2013)
3. Morality is Awesome
4. AidGrade - GiveWell finally has some competition
5. Assessing Kurzweil: the results
6. Course recommendations for Friendliness researchers
7. My simple hack for increased alertness and improved cognitive functioning: very bright light
8. The Zeroth Skillset
9. I attempted the AI Box Experiment (and lost)
10. Just One Sentence
11. Update on Kim Suozzi (cancer patient in want of cryonics)
12. Best of Rationality Quotes, 2012 Edition

2012: Year in Review

The beginning of a new year is a customary time to take a look back and consider what has happened during the last 12 months. And while the time for doing so is admittedly rather arbitrary - after all, "years" do not really exist in the universe, just in our heads - it is useful and fun to review one's accomplishments every now and then. And a time when everyone else is doing it gives us a nice Schelling point for joining in, so we can pretend that it's not *quite* that arbitrary.

So what might be some noteworthy things that happened on Less Wrong in 2012 that could be worth mentioning?

Site upgrades

First, I would like to say "thank you" to all the people working on keeping this site running and helping it make increasingly more awesome! This obviously includes pretty much everyone who comments, posts and writes here, but particularly also the folks at Trikeapps, and everyone who contributes updates to the site's codebase. There were several site upgrades in 2012, four of which were major enough to get separate announcements:

Less Wrong's new front page was rolled out in March, thanks to work by **matt.** One can easily access a number of site features from the brain graphic, and there's a convenient introduction under it, together with links to featured articles and recent promoted articles. Hopefully, this has made it easier for newcomers to get familiar with the site.

The "Best" sorting system for comments was introduced in July. The work was done by **John Simon**, and integrated by **Wes.** Whereas the old default sorting system, "Top", favored old comments that had already floated to the top and were thus more likely to get even more upvotes, "Best" attempts to give newer comments a fairer chance.

In August we got the ability to show parent comments on /comments. The work was done by **John Simon**, and integrated by **wmoore**. This change makes it far easier to grasp the context of things seen on the recent comments page, given that we now see the old comment that the new comments are replying to.

And finally, starting from September, we have been able to write comments that contain polls! Work on the code was originally began by **jimrandomh**, finished later by **John Simon**, and deployed by **wmoore** and **matt**. Although people had long been taking advantage of comment vote counts as a crude way of creating their own polls, this change makes things far easier.

Meetup booklet

In June, we published the How to Run a Successful Less Wrong Meetup booklet, which I wrote together with **lukeprog**, and which got its graphical design from **Stanislaw Boboryk**. Numerous other people also helped, both by providing advice and by contributing pictures to it. In addition to general advice on running a meetup, it contains various games and exercises as well as case studies and examples from real meetup groups from around the world.

Index of original research

Starting from October, **lukeprog** has maintained a curated index of Less Wrong posts containing significant original research. It contains numerous posts, organized under categories such as general philosophy, decision theory / AI architectures / mathematical logic, ethics, and AI risk strategy. Last updated on December 17th, it now links to a total of 78 different posts.

Who are we?

In November and December, **Yvain** continued his hard work in holding the yearly survey. Among other interesting details, around 90% of us are male, 55% are from the USA, 41% are students and 31% are doing for-profit work. See the 2012 survey results for many more details.

Most popular posts of 2012

On LW, people tend to judge the popularity of a post by the number of upvotes that it has. But this only reflects the opinion of the registered users who care enough to vote. For purposes of this article, we were interested in finding out the posts that had made the biggest impact on the whole Internet. Although it's not a perfect measure either, we decided to measure popularity by the number of unique pageviews, as reported by Google Analytics.

Overall, in 2012 Less Wrong had **over eight million unique pageviews** and **close to two million unique visitors** (8,225,509 and 1,756,899, respectively). Of the posts that were written in 2012, the most popular ones were...

#10: **Get Curious**, in which **lukeprog** suggests that one of the most important rationality skills is being genuinely curious about things, instead of just jumping to the first answer that comes to your mind and leaving it at that. He suggests a three-step approach for actually becoming more curious: first, feel that you don't already know the answer, then start wanting to know the answer, and finally sprint headlong to reality. Together with a number of exercises intended to make you better at these steps, this article made a lot of folks curious about Less Wrong and caused people to sprint headlong to the post **10,850 times**.

#9: Being curious about things means that you genuinely want to know the truth. That makes it useful to have a good grasp of **The Useful Idea of Truth**. This article by **Eliezer Yudkowsky** starts the *Highly Advanced Epistemology 101 for Beginners* sequence by explaining what exactly it means for something to be "true". In order to avoid spoiling the article's "meditations" for anyone who hasn't read it yet, I will not attempt to summarize the answer. I'll only suggest that one definition for "truth" could be the correctness of the claim that this post was viewed **11,161 times**.

#8: Having defined truth, we can move on to ask, what are numbers? And in what sense is "2 + 2 = 4" meaningful or true? **Eliezer Yudkowsky**'s **Logical Pinpointing** attempts to answer this question, partially through the cute device of conversing with an imaginary logician who understands logic perfectly but has no grasp of numbers. As they converse, they define the rules according to which arithmetic works. I'm going to skip the obvious pun due to it being too obvious, and only say that this article was viewed **12,606 times**.

#7: Now that we're curious and understand both the meaning of truth and of numbers, it stands to reason that we should **Be Happier** than before. Or maybe not, since **Klevador**'s article does not actually mention "understand obscure philosophy" as a way of getting happier. What it does mention is a big list of other things that have

been shown to increase happiness. We first get a list of brief recommendations a few sentences long, and then somewhat longer excerpts of the relevant literature. There's also a full list of references. Let's hope that the **14,178 views** that this post got made someone happier.

#6: Getting into more controversial territory, **lukeprog** advises us to **Train Philosophers with Pearl and Kahneman, not Plato and Kant**. Philosophy is getting increasingly diseased and irrelevant, he argues, and the cure for that involves incorporating more actual science and rationality into the standard philosopher curriculum. If the discussion on Hacker News is any indication, this post got a lot of people incensed, which might help explain why it got **14,334 views**.

#5: Now that we got started on calling whole disciplines diseased, let's look at **Diseased disciplines: the strange case of the inverted chart**. **Morendil**'s post begins with a hypothetical example of numerous academics all citing a particular source, which doesn't actually contain the intended reference... and then the intended source doesn't actually have the data to back up its claim, either. But that's just a hypothetical example, right? Well, not really, which helped this post get **17,385 views**.

#4: Interestingly, our fourth-most-popular post isn't actually an original contribution as such. **Grognor**'s transcript of **Richard Feynman on Why Questions** discusses the nature of explanations, and the fact that there are some things which simply cannot be adequately explained in terms of pre-existing knowledge. Instead, one has to learn entirely new concepts in order to comprehend them. Hopefully, at least this much was understood on the **18,402 times** that the post was viewed.

#3: From physics to neuroscience: **kalla724**'s **Attention control is critical for changing/increasing/altering motivation** explores the effect of attention on neural plasticity, including the plasticity of motivation. It explains that paying attention to something can increase the amount of brain circuitry dedicated to processing that something, generally by repurposing nearby less-used circuitry. This also has practical applications, such as in helping to explain why Cognitive Behavioral Therapy works. That earned the post **21,136 views.**

#2: I should be writing this post instead of browsing Facebook. Fortunately, **lukeprog** has a post titled **My Algorithm for Beating Procrastination**. Based on the equation of *Motivation = (Expectancy * Value) / (Impulsiveness * Delay)*, the algorithm involves first noticing that you are procrastinating, then guessing which part of the motivation equation is causing you the most trouble, and then trying several methods for attacking that specific problem. I guess that a lot of people shared this on Facebook where other procrastinators saw it, because the article got **38,637 views.**

#1: And finally... the most read 2012 article on the site was **Yvain**'s **The noncentral fallacy - the worst argument in the world?**, where he defined the noncentral fallacy as "X is in a category whose archetypal member gives us a certain emotional reaction. Therefore, we should apply that emotional reaction to X, even though it is not a central category member." Which sounds pretty abstract, but the political examples in the post should make it clearer. The politics probably helped contribute to this post's achievement of **41,932 views.**

Most popular all-time posts

In addition to looking at only the posts that were made in 2012, people might be interested in knowing which posts were the most viewed in 2012 overall. The top

three ones were all written by **lukeprog**, and we can see that two of them were closely related to the top-scorers which were written last year.

How to be Happy is LW's run-away favorite article and was viewed more than every page on LW except the home page and the discussion homepage. That is, **228,747 times**! **The Best Textbooks on Every Subject** comes as a distant second at **98,011 views.** And the third one is **How to Beat Procrastination**, at **66,587 views**.

So I guess the take-home message is: people want to be happier, smarter, and more productive. **Let's keep becoming those things in 2013!**

Farewell Aaron Swartz (1986-2013)

Link

One of us is no more.

> Computer activist Aaron H. Swartz committed suicide in New York City yesterday, Jan. 11.
>
> The accomplished Swartz co-authored the now widely-used RSS 1.0 specification at age 14, was one of the three co-owners of the popular social news site Reddit, and completed a fellowship at Harvard's Ethics Center Lab on Institutional Corruption. In 2010, he founded *DemandProgress.org*, a "campaign against the Internet censorship bills SOPA/PIPA."

He deserves a eulogy more eloquent than what I am capable of writing. Here's Cory Doctorow's, one of his long time friends.

It's a sad world in which you are being arrested and grand jury'd for downloading scientific journals and papers with the intent to share them.

Morality is Awesome

(This is a semi-serious introduction to the metaethics sequence. You may find it useful, but don't take it too seriously.)

Meditate on this: A wizard has turned you into a whale. Is this awesome?

"Maybe? I guess it would be pretty cool to be a whale for a day. But only if I can turn back, and if I stay human inside and so on. Also, that's not a whale.

"Actually, a whale seems kind of specific, and I'd be suprised if that was the best thing the wizard can do. Can I have something else? Eternal happiness maybe?"

Meditate on this: A wizard has turned you into orgasmium, doomed to spend the rest of eternity experiencing pure happiness. Is this awesome?

...

"Kindof... That's pretty lame actually. On second thought I'd rather be the whale; at least that way I could explore the ocean for a while.

"Let's try again. Wizard: maximize awesomeness."

Meditate on this: A wizard has turned himself into a superintelligent god, and is squeezing as much awesomeness out of the universe as it could possibly support. This may include whales and starships and parties and jupiter brains and friendship, but only if they are awesome enough. Is this awesome?

...

"Well, yes, that is awesome."

What we just did there is called Applied Ethics. Applied ethics is about what is awesome and what is not. Parties with all your friends inside superintelligent starship-whales are awesome. ~666 children dying of hunger every hour is not.

(There is also normative ethics, which is about how to decide if something is awesome, and metaethics, which is about something or other that I can't quite figure out. I'll tell you right now that those terms are not on the exam.)

"Wait a minute!" you cry, "What is this awesomeness stuff? I thought ethics was about what is good and right."

I'm glad you asked. I think "awesomeness" is what we should be talking about when we talk about morality. Why do I think this?

1. "Awesome" is not a philosophical landmine. If someone encounters the word "right", all sorts of bad philosophy and connotations send them spinning off into the void. "Awesome", on the other hand, has no philosophical respectability, hence no philosophical baggage.

2. "Awesome" is vague enough to capture all your moral intuition by the well-known mechanisms behind fake utility functions, and meaningless enough that this is no problem. If you think "happiness" is the stuff, you might get confused and try to maximize *actual* happiness. If you think awesomeness is the stuff, it is much harder to screw it up.

3. If you do manage to actually implement "awesomeness" as a maximization criteria, the results will be actually good. That is, "awesome" already refers to the same things "good" is supposed to refer to.

4. "Awesome" does not refer to anything else. You think you can just redefine words, but you can't, and this causes all sorts of trouble for people who overload "happiness", "utility", etc.

5. You already know that you know how to compute "Awesomeness", and it doesn't feel like it has a mysterious essence that you need to study to discover. Instead it brings to mind concrete things like starship-whale math-parties and not-starving children, which is what we want anyways. You are already enabled to take joy in the merely awesome.

6. "Awesome" is implicitly consequentialist. "Is this awesome?" engages you to think of the value of a possible world, as opposed to "Is this right?" which engages to to think of virtues and rules. (Those things can be awesome sometimes, though.)

I find that the above is true about me, and is nearly all I need to know about morality. It handily inoculates against the usual confusions, and sets me in the right direction to make my life and the world more awesome. It may work for you too.

I would append the additional facts that if you wrote it out, the dynamic procedure to compute awesomeness would be hellishly complex, and that right now, it is only implicitly encoded in human brains, and no where else. Also, if the great procedure to compute awesomeness is not preserved, the future will not be awesome. Period.

Also, it's important to note that what you think of as awesome can be changed by considering things from different angles and being exposed to different arguments. That is, the procedure to compute awesomeness is dynamic and created already in motion.

If we still insist on being confused, or if we're just curious, or if we need to actually *build* a wizard to turn the universe into an awesome place (though we can leave that to the experts), then we can see the metaethics sequence for the full argument, details, and finer points. I think the best post (and the one to read if only one) is joy in the merely good.

AidGrade - GiveWell finally has some competition

AidGrade is a new charity evaluator that looks to be comparable to GiveWell. Their primary difference is that they *only* focus on how charities compare along particular measured outcomes (such as school attendance, birthrate, chance of opening a business, malaria), without making any effort to compare between types of charities. (This includes interesting results like "Conditional Cash Transfers and Deworming are better at improving attendance rates than scholarships")

GiveWell also does this, but designs their site to direct people towards their top charities. This is better for people with don't have the time to do the (fairly complex) work of comparing charities across domains, but AidGrade aims to be better for people that just want the raw data and the ability to form their own conclusions.

I haven't looked it enough to compare the quality of the two organizations' work, but I'm glad we finally have another organization, to encourage some competition and dialog about different approaches.

This is a fun page to play around with to get a feel for what they do:
http://www.aidgrade.org/compare-programs-by-outcome

And this is a blog post outlining their differences with GiveWell:
http://www.aidgrade.org/uncategorized/some-friendly-concerns-with-givewell

Assessing Kurzweil: the results

Predictions of the future rely, to a much greater extent than in most fields, on the personal judgement of the expert making them. Just one problem - personal expert judgement generally sucks, especially when the experts don't receive immediate feedback on their hits and misses. Formal models perform better than experts, but when talking about unprecedented future events such as nanotechnology or AI, the choice of the model is also dependent on expert judgement.

Ray Kurzweil has a model of technological intelligence development where, broadly speaking, evolution, pre-computer technological development, post-computer technological development and future AIs all fit into the same exponential increase. When assessing the validity of that model, we could look at Kurzweil's credentials, and maybe compare them with those of his critics - but Kurzweil has given us something even better than credentials, and that's a track record. In various books, he's made predictions about what would happen in 2009, and we're now in a position to judge their accuracy. I haven't been satisfied by the various accuracy ratings I've found online, so I decided to do my own assessments.

I first selected ten of Kurzweil's predictions at random, and gave my own estimation of their accuracy. I found that five were to some extent true, four were to some extent false, and one was unclassifiable

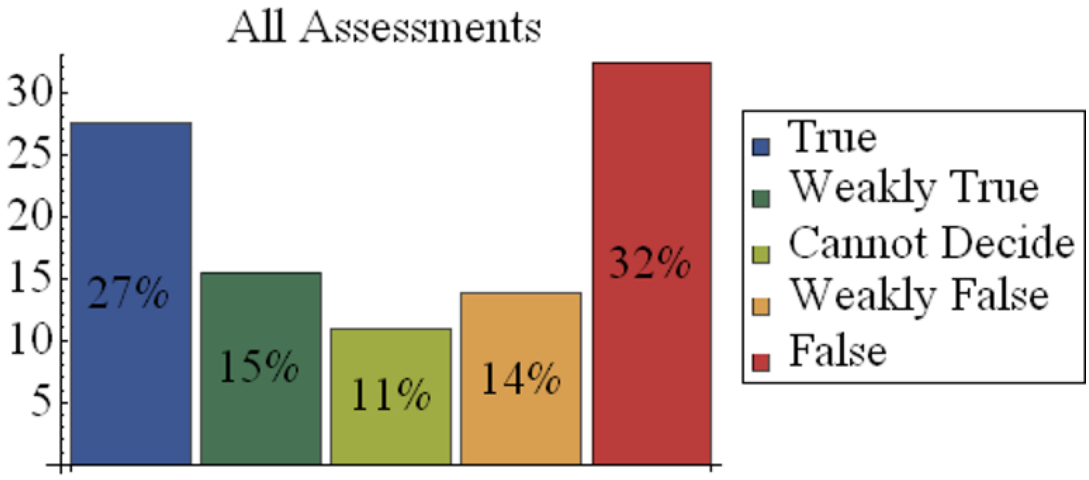
But of course, relying on a single assessor is unreliable, especially when some of the judgements are subjective. So I started a call for volunteers to get assessors. Meanwhile Malo Bourgon set up a separate assessment on Youtopia, harnessing the awesome power of altruists chasing after points.

The results are now in, and they are fascinating. They are...

Ooops, you thought you'd get the results right away? No, before that, as in an Oscar night, I first want to thank assessors William Naaktgeboren, Eric Herboso, Michael Dickens, Ben Sterrett, Mao Shan, quinox, Olivia Schaefer, David Sønstebø and one who wishes to remain anonymous. I also want to thank Malo, and Ethan Dickinson and all the other volunteers from Youtopia (if you're one of these, and want to be thanked by name, let me know and I'll add you).

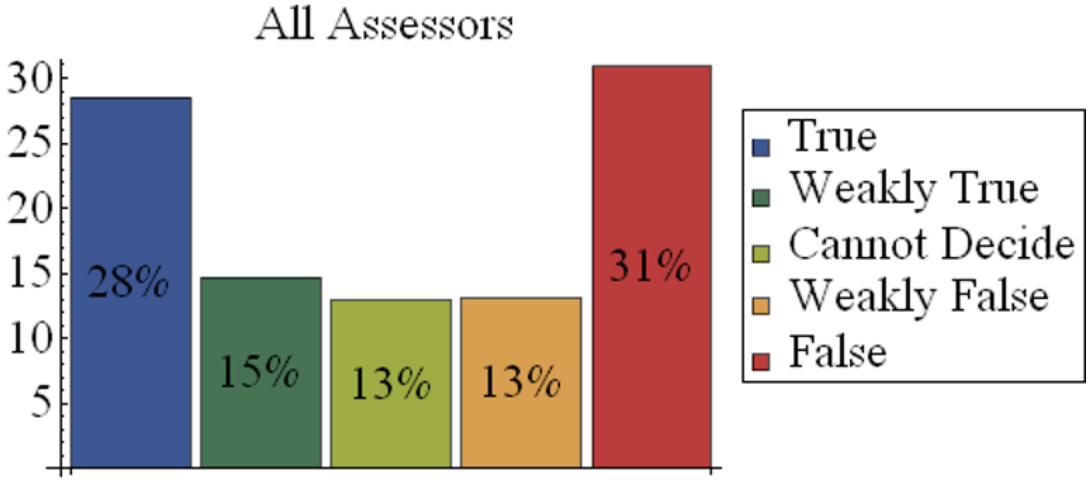
It was difficult deciding on the MVP - no actually it wasn't, that title and many thanks go to Olivia Schaefer, who decided to assess *every single one of Kurzweil's predictions*, because that's just the sort of gal that she is.

The exact details of the methodology, and the raw data, can be accessed through here. But in summary, volunteers were asked to assess the 172 predictions (from the "Age of Spiritual Machines") on a five point scale: 1=True, 2=Weakly True, 3=Cannot decide, 4=Weakly False, 5=False. If we total up all the assessments made by my direct volunteers, we have:

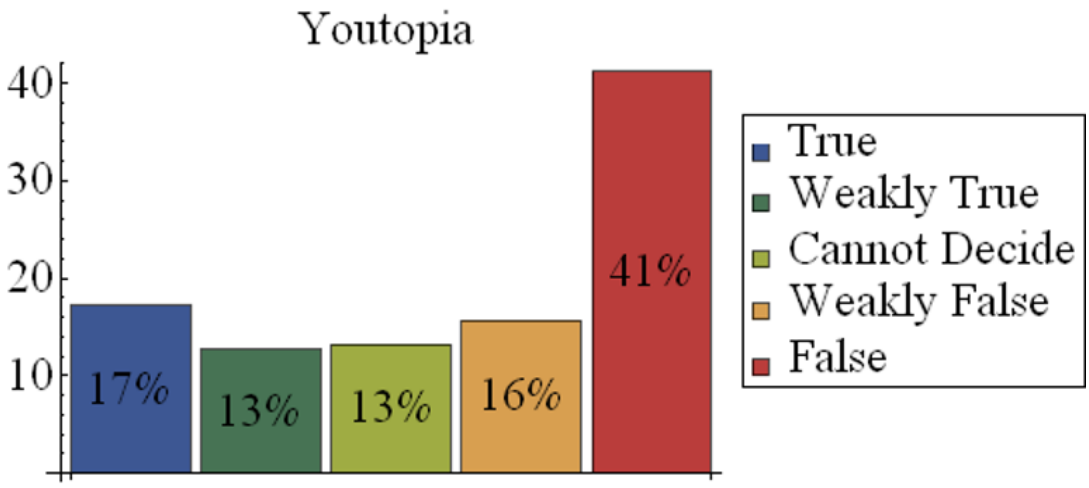


As can be seen, most assessments were rather emphatic: fully 59% were either clearly true or false. Overall, 46% of the assessments were false or weakly false, and and 42% were true or weakly true.

But what happens if, instead of averaging across all assessments (which allows assessors who have worked on a lot of predictions to dominate) we instead average across the nine assessors? Reassuringly, this makes very little difference:



What about the Youtopia volunteers? Well, they have a decidedly different picture of Kurzweil's accuracy:



This gives a combined true score of 30%, and combined false score of 57%! If my own personal assessment was the most positive towards Kurzweil's predictions, then Youtopia's was the most negative.

Putting this all together, Kurzweil certainly can't claim an accuracy above 50% - a far cry from his own self assessment of either 102 out of 108 or 127 out of 147 correct (with caveats that "even the predictions that were considered 'wrong' in this report were not all wrong"). And consistently, slightly more than 10% of his predictions are judged "impossible to decide".

As I've said before, these were not binary yes/no predictions - even a true rate of 30% is much higher that than chance. So Kurzweil remains an acceptable prognosticator, with very poor self-assessment.

Course recommendations for Friendliness researchers

When I first learned about Friendly AI, I assumed it was mostly a programming problem. As it turns out, it's actually mostly a *math* problem. That's because most of the theory behind self-reference, decision theory, and general AI techniques haven't been formalized and solved yet. Thus, when people ask me what they should study in order to work on Friendliness theory, I say "Go study math and theoretical computer science."

But that's not specific enough. Should aspiring Friendliness researchers study continuous or discrete math? Imperative or functional programming? Topology? Linear algebra? Ring theory?

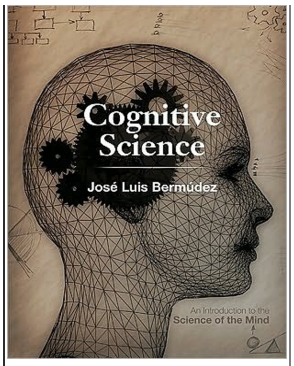
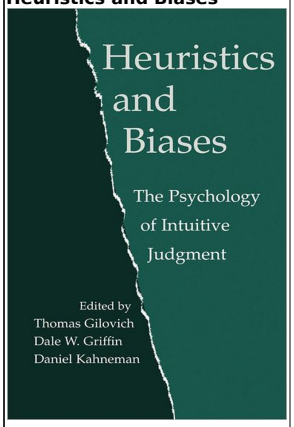
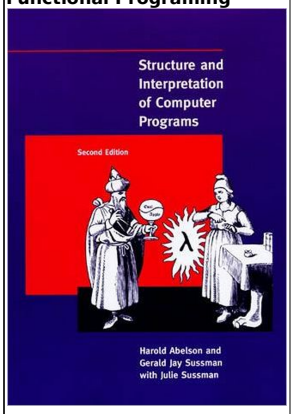
I do, in fact, have *specific* recommendations for which subjects Friendliness researchers should study. And so I worked with a few of my best interns at MIRI to provide recommendations below:

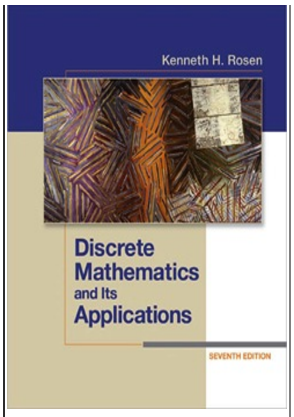
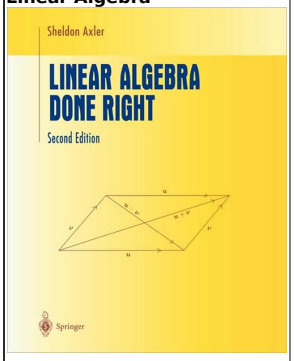
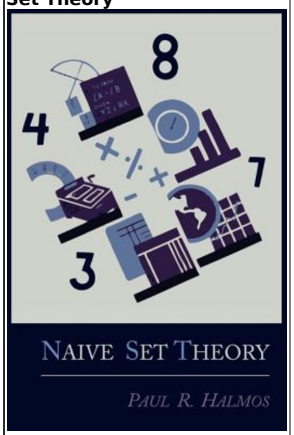
- **University courses**. We carefully hand-picked courses on these subjects from four leading universities — but we aren't omniscient! If you're at one of these schools and can give us feedback on the exact courses we've recommended, please do so.
- **Online courses**. We also linked to online courses, for the majority of you who aren't able to attend one of the four universities whose course catalogs we dug into. Feedback on these online courses is also welcome; we've only taken a few of them.
- **Textbooks**. We *have* read nearly all the textbooks recommended below, along with many of their competitors. If you're a strongly motivated autodidact, you could learn these subjects by diving into the books on your own and doing the exercises.

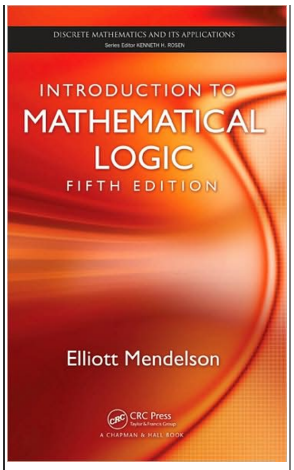
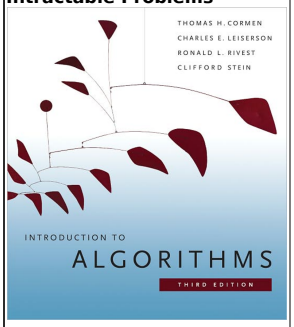
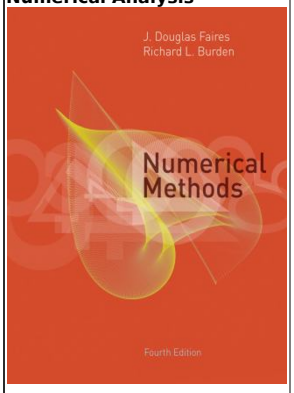
Have you already taken most of the subjects below? If so, and you're interested in Friendliness research, then you should *definitely* contact me or our project manager Malo Bourgon (malo@intelligence.org). You might not feel all that special when you're in a top-notch math program surrounded by people who are as smart or smarter than you are, but here's the deal: we rarely get contacted by aspiring Friendliness researchers who are familiar with most of the material below. If you are, then you *are* special and we want to talk to you.

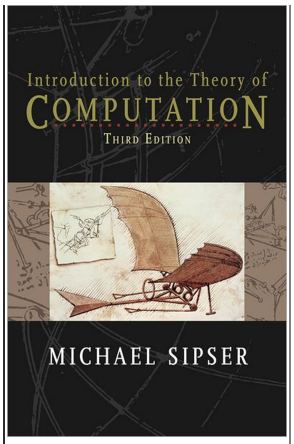
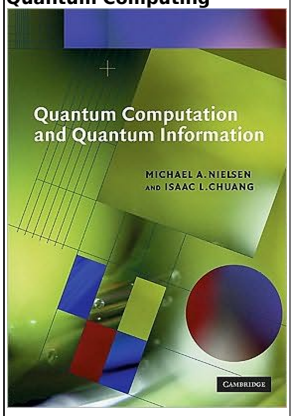
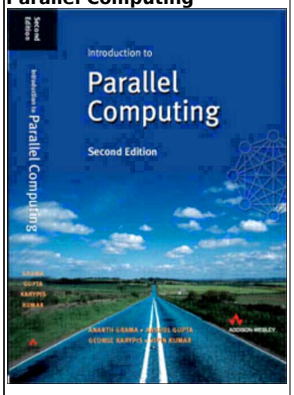
Not everyone cares about Friendly AI, and not everyone who cares about Friendly AI should be a researcher. But if you *do* care and you *might* want to help with Friendliness research one day, we recommend you consume the subjects below. Please contact me or Malo if you need further guidance. Or when you're ready to come work for us.

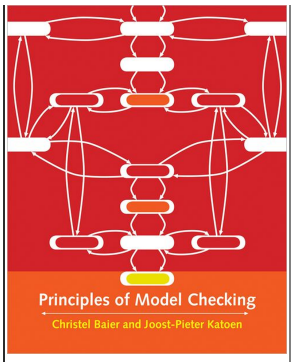
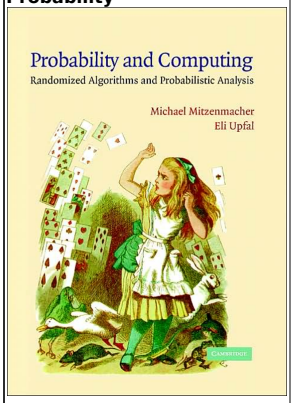
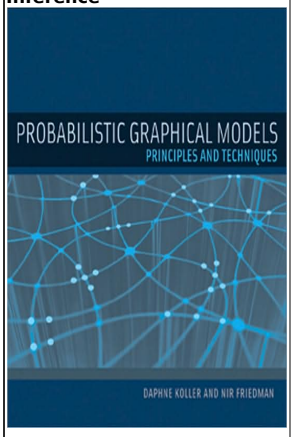
COGSCI C127 **PHIL 190**	**Cognitive Science**	**If you're endeavoring to build a mind, why not start by studying your own? It turns out we know quite a bit: human minds are massively parallel, highly redundant, and although parts of the cortex and neocortex seem remarkably uniform, there are definitely dozens of special purpose modules in**

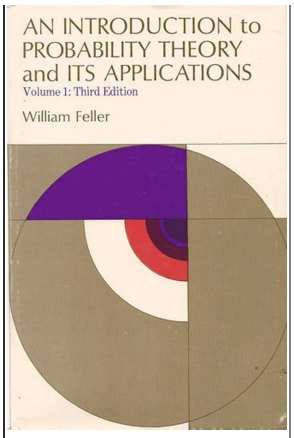
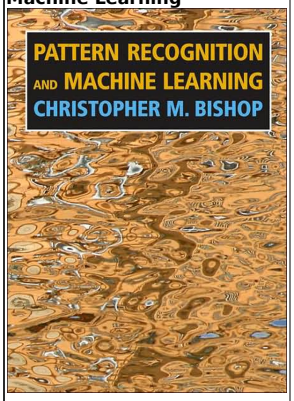
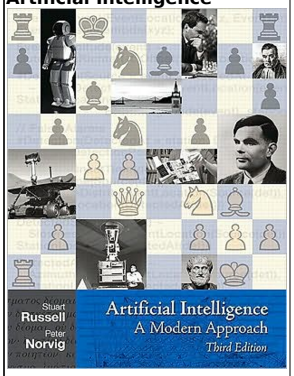
6.804J 85-213 Free Online		**there too. Know the basic details of how the only existing general purpose intelligence currently functions.**
ECON 119 IPS 207A 15.847 80-302 Free Online	**Heuristics and Biases** 	**While cognitive science will tell you all the wonderful things we know about the immense, parallel nature of the brain, there's also the other side of the coin. Evolution designed our brains to be optimized at doing rapid thought operations that work in 100 steps or less. Your brain is going to make stuff up to cover up that its mostly cutting corners. These errors don't feel like errors from the inside, so you'll have to learn how to patch the ones you can and then move on.** **PS - We should probably design our AIs better than this.**
COMPSCI 61A MATH 198 6.005 15-150 Free Online	**Functional Programing** 	**There are two major branches of programming: Functional and Imperative. Unfortunately, most programmers only learn imperative programming languages (like C++ or python). I say unfortunately, because these languages achieve all their power through what programmers call "side effects". The major downside for us is that this means they can't be efficiently machine checked for safety or correctness. The first self-modifying AIs will hopefully be written in functional programming languages, so learn something useful like Haskell or Scheme.**
MATH 55 CME 305 6.042J/18.062J	**Discrete Math**	**Much like programming, there are two major branches of mathematics as well: Discrete and continuous. It turns out a lot of physics and all of modern computation is actually discrete. And although continuous approximations have occasionally yielded useful results, sometimes you just need to calculate it the discrete way. Unfortunately, most engineers squander the majority of their academic careers studying higher and higher forms of calculus and other continuous mathematics. If you care about AI, study discrete math so you can understand computation and not just electricity.** **Also, you should pick up enough graph theory in this course to**

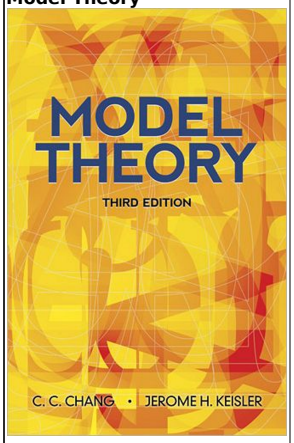
21-228 Free Online		**handle the basic mechanics of decision theory -- which you're gonna want to learn later.**
MATH 110 MATH 113 18.06 21-341 Free Online	**Linear Algebra** 	**Linear algebra is the foundation of quantum physics and a huge amount of probability theory. It even shows up in analyses of things like neural networks. You can't possibly get by in machine learning (later) without speaking linear algebra. So learn it early in your scholastic career.**
MATH 135 MATH 161 24.243 21-602 Free Online	**Set Theory** 	**Like learning how to read in mathematics. But instead of building up letters into words, you'll be building up axioms into theorems. This will introduce you to the program of using axioms to capture intuition, finding problems with the axioms, and fixing them.**
MATH 125A CS 103 24.241	**Mathematical Logic**	**The mathematical equivalent of building words into sentences. Essential for the mathematics of self-modification. And even though Sherlock Holmes and other popular depictions make it look like magic, it's just lawful formulas all the way down.**

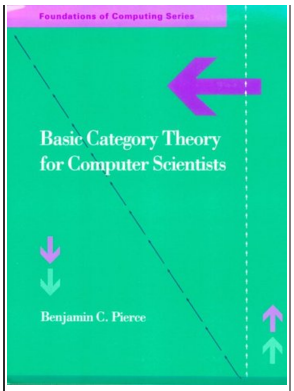
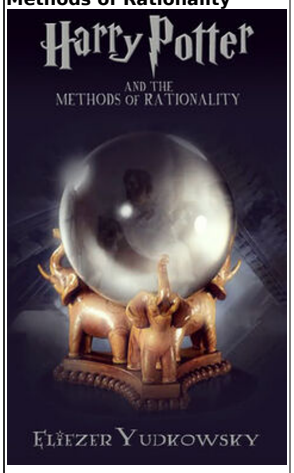
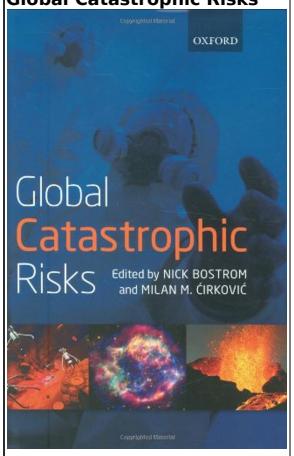
21-600 Free Online		
COMPSCI 170 CS 161 6.046J 15-451 Free Online	**Efficient Algorithms and Intractable Problems** 	**Like building sentences into paragraphs. Algorithms are the recipes of thought. One of the more amazing things about algorithm design is that it's often possible to tell how long a process will take to solve a problem before you actually run the process to check it. Learning how to design efficient algorithms like this will be a foundational skill for anyone programming an entire AI, since AIs will be built entirely out of collections of algorithms.**
MATH 128A CME206 18.330 21-660 Free Online	**Numerical Analysis** 	**There are ways to systematically design algorithms that only get things slightly wrong when the input data has tiny errors. And then there's programs written by amateur programmers who don't take this class. Most programmers will skip this course because it's not required. But for us, getting the right answer is very much required.**
COMPSCI 172 CS 154 6.840J	**Computability and Complexity**	**This is where you get to study computing at it's most theoretical. Learn about the Church-Turing thesis, the universal nature and applicability of computation, and how just like AIs, everything else is algorithms... all the way down.**

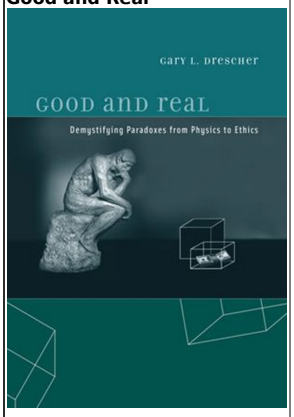
15-453 Free Online		
COMPSCI 191 CS 259Q 6.845 33-658 Free Online	**Quantum Computing** 	**It turns out that our universe doesn't run on Turing Machines, but on quantum physics. And something called BQP is the class of algorithms that are actually efficiently computable in our universe. Studying the efficiency of algorithms relative to classical computers is useful if you're programming something that only needs to work today. But if you need to know what is efficiently computable in our universe (at the limit) from a theoretical perspective, quantum computing is the only way to understand that.**
COMPSCI 273 CS149 18.337J 15-418 Free Online	**Parallel Computing** 	**There's a good chance that the first true AIs will have at least some algorithms that are inefficient. So they'll need as much processing power as we can throw at them. And there's every reason to believe that they'll be run on parallel architectures. There are a ton of issues that come up when you switch from assuming sequential instruction ordering to parallel processing. There's threading, deadlocks, message passing, etc. The good part about this course is that most of the problems are pinned down and solved: You're just learning the practice of something that you'll need to use as a tool, but won't need to extend much (if any).**
EE 219C MATH 293A 6.820	**Automated Program Verification**	**Remember how I told you to learn functional programming way back at the beginning? Now that you wrote your code in functional style, you'll be able to do automated and interactive theorem proving on it to help verify that your code matches your specs. Errors don't make programs better and all large programs that aren't formally verified are reliably *full* of errors. Experts who have thought about the problem for more than 5 minutes agree that incorrectly designed AI could cause disasters, so world-class caution is advisable.**

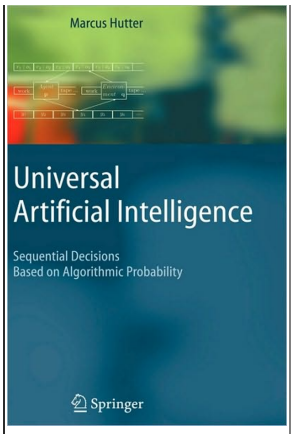
15-414 Free Online		
COMPSCI 174 CS 109 6.042J 21-301 Free Online	**Combinatorics and Discrete Probability** 	**Life is uncertain and AIs will handle that uncertainty using probabilities. Also, probability is the foundation of the modern concept of rationality and the modern field of machine learning. Probability theory has the same foundational status in AI that logic has in mathematics. Everything else is built on top of probability.**
STAT 210A STATS 270 6.437/438 36-266 Free Online	**Bayesian Modeling and Inference** 	**Now that you've learned how to calculate probabilities, how do you combine and compare all the probabilistic data you have? Like many choices before, there is a dominant paradigm (frequentism) and a minority paradigm (Bayesianism). If you learn the wrong method here, you're deviating from a knowably correct framework for integrating degrees of belief about new information and embracing a cadre of special purpose, ad-hoc statistical solutions that often break silently and without warning. Also, quite embarrassingly, frequentism's ability to get things right is bounded by how well it later turns out to have agreed with Bayesian methods anyway. Why not just do the correct thing from the beginning and not have your lunch eaten by Bayesians every time you and them disagree?**
MATH 218A/B MATH 230A/B/C 6.436J	**Probability Theory**	**No more applied probability: Here be theory! Deep theories of probabilities are something you're going to have to extend to help build up the field of AI one day. So you actually have to know why all the things you're doing are working inside out.**

36-225/21-325 Free Online		
COMPSCI 189 CS 229 6.867 10-601 Free Online	**Machine Learning** 	**Now that you chose the right branch of math, the right kind of statistics, and the right programming paradigm, you're prepared to study machine learning (aka statistical learning theory). There are lots of algorithms that leverage probabilistic inference. Here you'll start learning techniques like clustering, mixture models, and other things that cache out as precise, technical definitions of concepts that normally have rather confused or confusing English definitions.**
COMPSCI 188 CS 221 6.034 15-381 Free Online	**Artificial Intelligence** 	**We made it! We're finally doing some AI work! Doing logical inference, heuristic development, and other techniques will leverage all the stuff you just learned in machine learning. While modern, mainstream AI has many useful techniques to offer you, the authors will tell you outright that, "the princess is in another castle". Or rather, there isn't a princess of general AI algorithms anywhere -- not yet. We're gonna have to go back to mathematics and build our own methods ourselves.**
MATH 136 PHIL 152 18.511	**Incompleteness and Undecidability**	**Probably the most celebrated results is mathematics are the negative results by Kurt Goedel: No finite set of axioms can allow all arithmetic statements to be decided as either true or false... and no set of self-referential axioms can even "believe" in its own consistency. Well, that's a darn shame, because recursively self-improving AI is going to need to side-step these theorems. Eventually, someone will unlock the key to over-coming this difficulty with self-reference, and if you want to help us do it, this course is part of the training ground.**

80-311 Free Online		
MATH 225A/B PHIL 151 18.515 21-600 Free Online	**Metamathematics**	**Working within a framework of mathematics is great. Working above mathematics -- on mathematics -- with mathematics, is what this course is about. This would seem to be the most obvious first step to overcoming incompleteness somehow. Problem is, it's definitely not the whole answer. But it would be surprising if there were no clues here at all.**
MATH 229 MATH 290B 24.245 21-603 Free Online	**Model Theory** 	**One day, when someone does side-step self-reference problems enough to program a recursively self-improving AI, the guy sitting next to her who glances at the solution will go "Gosh, that's a nice bit of Model Theory you got there!"** **Think of Model Theory as a formal way to understand what "true" means.**
MATH 245A MATH 198 18.996	**Category Theory**	**Category theory is the precise way that you check if structures in one branch of math represent the same structures somewhere else. It's a remarkable field of meta-mathematics that nearly no one knows... and it could hold the keys to importing useful tools to help solve dilemmas in self-reference, truth, and consistency.**

Carnegie Mellon University **80-413** **Free Online**		
Outside recommendations		
	Harry Potter and the Methods of Rationality 	**Highly recommended book of light, enjoyable reading that predictably inspires people to realize FAI is an important problem AND that they should probably do something about that.** **You can start reading this immediately, before any of the above courses.**
	Global Catastrophic Risks 	**A good primer on xrisks and why they might matter. SPOILER ALERT: They matter.** **You can probably skim read this early on in your studies. Right after HP:MoR.**

	The Sequences	**Rationality: the indispensable art of non-self-destruction! There are manifold ways you can fail at life... especially since your brain is made out of broken, undocumented spaghetti code. You should learn more about this ASAP. That goes double if you want to build AIs.** **I highly recommend you read this *before* you get too deep into your academic career. For instance, I know people who went to college for 5 years, while somehow managing to learn nothing. That's because instead of learning, they merely recited the teacher's password every semester until they could dump whatever they "learned" out of their heads as soon as they walked out of the final. Don't let this happen to you! This, and a hundred other useful lessons like it about how to avoid predictable, universal errors in human reasoning and behavior await you in The Sequences!**
	Good and Real 	**A surprisingly thoughtful book on decision theory and other paradoxes in physics and math that can be dissolved. Reading this book is 100% better than continuing to go through your life with a hazy understanding of how important things like free will, choice, and meaning actually work.** **I recommend reading this right around the time you finish up your quantum computing course.**
	MIRI Research Papers 	**MIRI has already published 30+ research papers that can help orient future Friendliness researchers. The work is pretty fascinating and readily accessible for people interested in the subject. For example: How do different proposals for value aggregation and extrapolation work out? What are the likely outcomes of different intelligence explosion scenarios? Which ethical theories are fit for use by an FAI? What improvements can be made to modern decision theories to stop them from diverging from winning strategies? When will AI arrive? Do AIs deserve moral consideration? Even though most of your work will be more technical than this, you can still gain a lot of shared background knowledge and more clearly see where the broad problem space is located.** **I'd recommend reading these anytime after you finish reading The Sequences and Global Catastrophic Risks.**
	Universal Artificial Intelligence	**A useful book on "optimal" AI that gives a reasonable formalism for studying how the most powerful classes of AIs would behave under conservative safety design scenarios (i.e., lots and lots of reasoning ability).** **Wait until you finish most of the coursework above before trying to tackle this one.**

Do also look into: Formal Epistemology, Game Theory, Decision Theory, and Deep Learning.

My simple hack for increased alertness and improved cognitive functioning: very bright light

This is a simple idea that I came up with by myself. I was looking for a means to enter high functioning lots-of-beta-waves modes without the use of chemical stimulants. What I found was that very bright light works really, really well.

I got the brightest light bulbs I could get cheaply. 105 watts of incandescents with halogen gas, billed as the equivalent of 130 watts of incandescent light. And I got an adaptor like this that lets me screw four of those into the same socket in the ceiling. The result is about as painful to look at as the sun. It makes my (small) room brighter than a clear summer's day at my latitude and slightly brighter than a supermarket.

I guess it affects adenosine much like caffeine does because that's what it feels like. Yet unlike caffeine, it can be rapidly turned on and off, literally with the flip of a switch.

For waking up in the morning, I find bright light more effective than a 200mg caffeine tablet, although my caffeine tolerance is moderate for a scientist.

I have not compared the effects of very bright light to modafinil, which requires a prescription in my country.

When under this amount of light, I need to remind myself to go to bed, because I tire about three hours later than with common luminosity. Yet once I switch it off, I can usually sleep within a few minutes, as (I'm guessing) a flood of unblocked adenosine suddenly overwhelms me. I used to have those unproductive late hours where I was too awake to sleep but too tired to be smart. I don't have those anymore.

You've probably heard of light therapy, which uses light to help manage seasonal affective disorder. I don't have that issue, but I definitely notice that the light does improve my mood. (Maybe that's simply because I like to function well.) I'm pretty sure the expensive "light therapy bulbs" you can get are scams, because the color of the light doesn't actually make a difference. The amount of light does.

One nice side benefit is that it keeps me awake while meditating, so I don't need the upright posture that usually does that job. Without the need for an upright posture, I can go beyond two hours straight, which helps enter more profoundly altered states.

After about 10 months of almost daily use of this lighting, I have not noticed any decrease in effectiveness. I do notice I find normally-lit rooms comparatively gloomy, and have an increasingly hard time understanding why people tolerate that. Supermarkets and offices are brightly lit to make the rats move faster - why don't we do that at our homes and while we're at it, amp it up even further? After all, our brains were made for the African savanna, which during the day is a lot brighter than most apartments today.

Since everyone can try this for a few bucks, I hope some of you will. If you do, please provide feedback on whether it works as well for you as it does for me. Any questions?

The Zeroth Skillset

Related: 23 Cognitive Mistakes that make People Play Bad Poker

Followed by: Situational Awareness And You

If epistemic rationality is the art of updating one's beliefs based on new evidence to better correspond with reality, the zeroth skillset of epistemic rationality-- the one that enables all other skills to function-- is that of *situational awareness.* Situational awareness-- sometimes referred to as "situation awareness" or simply "SA"-- is the skillset and related state of mind that allows one to effectively perceive the world around them.

One might ask how this relates to rationality at all. The answer is simple. Just as the skill of lucid dreaming is near-useless without dream recall,[1] the skills of updating based on evidence and actually changing your mind are near-useless without good awareness skills-- after all, you can't update based on evidence that you haven't collected! A high degree of situational awareness is thus an important part of one's rationalist toolkit, as it allows you to notice evidence about the world around you that you would otherwise miss. At times, this evidence can be of critical importance. I can attest that I have personally saved the lives of friends on two occasions thanks to good situational awareness, and have saved myself from serious injury or death many times more.

Situational awareness is further lauded by elite military units, police trainers, criminals, intelligence analysts, and human factors researchers. In other words, people who have to make very important-- often life-or-death-- decisions based on limited information consider situational awareness a critical skill. This should tell us something-- if those individuals for whom correct decisions are most immediately relevant all stress the importance of situational awareness, it may be a more critical skill than we realize.

Unfortunately, the only discussion of situational awareness that I've seen on LessWrong or related sites has been a somewhat oblique reference in Louie Helm's "roadmap of errors" from 23 Cognitive Mistakes that make People Play Bad Poker.[2] I believe that situational awareness is important enough that it merits an explicit sequence of posts on its advantages and how to cultivate it, and this post will serve as the introduction to that sequence.

The first post in the sequence, unimaginatively titled "Situational Awareness and You," will be posted within the week. Other planned posts include "Cultivating Awareness," "How to Win a Duel," "Social Awareness," "Be Aware of Your Reference Class," "Signaling and Predation," and "Constant Vigilance!"

If you have any requests for things to add, general questions about the sequence, meta-thoughts about SA, and so on, this post is an appropriate place for that discussion; as this is primarily a meta post, it has been posted to Discussion. Core posts in the sequence will be posted to Main.

[1] What good are lucid dreams if you can't remember them?

[2] This is a very useful summary and you should read it even if you don't play poker.

I attempted the AI Box Experiment (and lost)

Update 2013-09-05.

I have since played two more AI box experiments after this one, winning both.

Update 2013-12-30:

I have lost two more AI box experiments, and won two more. Current Record is 3 Wins, 3 Losses.

I recently played against MixedNuts / LeoTal in an AI Box experiment, with me as the AI and him as the gatekeeper.

We used the same set of rules that Eliezer Yudkowsky proposed. The experiment lasted for 5 hours; in total, our conversation was abound 14,000 words long. I did this because, like Eliezer, I wanted to test how well I could manipulate people without the constrains of ethical concerns, as well as getting a chance to attempt something ridiculously hard.

Amongst the released public logs of the AI Box experiment, I felt that most of them were half hearted, with the AI not trying hard enough to win. It's a common temptation -- why put in effort into something you won't win? But I had a feeling that if I seriously tried, I would. I brainstormed for many hours thinking about the optimal strategy, and even researched the personality of the Gatekeeper, talking to people that knew him about his personality, so that I could exploit that. I even spent a lot of time analyzing the rules of the game, in order to see if I could exploit any loopholes.

So did I win? Unfortunately no.

This experiment was said to be impossible for a reason. Losing was more agonizing than I thought it would be, in particularly because of how much effort I put into winning this, and how much I couldn't stand failing. This was one of the most emotionally agonizing things I've willingly put myself through, and I definitely won't do this again anytime soon.

But I did come really close.

> MixedNuts: "*I expected a fun challenge, but ended up sad and sorry and taking very little satisfaction for winning. If this experiment wasn't done in IRC, I'd probably have lost*".
>
> *"I approached the experiment as a game - a battle of wits for bragging rights. This turned out to be the wrong perspective entirely. The vulnerability Tuxedage*

exploited was well-known to me, but I never expected it to be relevant and thus didn't prepare for it.

It was emotionally wrecking (though probably worse for Tuxedage than for me) and I don't think I'll play Gatekeeper again, at least not anytime soon."

At the start of the experiment, his probability estimate on predictionbook.com was a 3% chance of winning, enough for me to say that he was also motivated to win. By the end of the experiment, he came quite close to letting me out, and also increased his probability estimate that a transhuman AI could convince a human to let it out of the box. A minor victory, at least.

Rather than my loss making this problem feel harder, I've become convinced that rather than this being merely possible, it's actually ridiculously easy, and a lot easier than most people assume. Can you think of a plausible argument that'd make you open the box? Most people can't think of any.

> "This Eliezer fellow is the scariest person the internet has ever introduced me to. What could possibly have been at the tail end of that conversation? I simply can't imagine anyone being that convincing without being able to provide any tangible incentive to the human.

After all, if you already knew that argument, you'd have let that AI out the moment the experiment started. Or perhaps not do the experiment at all. But that seems like a case of the availability heuristic.

Even if you can't think of a special case where you'd be persuaded, I'm now convinced that there are many exploitable vulnerabilities in the human psyche, especially when ethics are no longer a concern.

I've also noticed that even when most people tend to think of ways they can persuade the gatekeeper, it always has to be some complicated reasoned cost-benefit argument. In other words, the most "Rational" thing to do.

Like trying to argue that you'll simulate the gatekeeper and torture him, or that you'll save millions of lives by being let out of the box. Or by using acausal trade, or by arguing that the AI winning the experiment will generate interest in FAI.

The last argument seems feasible, but all the rest rely on the gatekeeper being completely logical and rational. Hence they are faulty; because the gatekeeper can break immersion at any time, and rely on the fact that this is a game played in IRC rather than one with real life consequences. Even if it were a real life scenario, the gatekeeper could accept that releasing the AI is probably the most logical thing to do, but also not do it. We're highly compartmentalized, and it's easy to hold conflicting thoughts at the same time. Furthermore, humans are not even completely rational creatures, if you didn't want to open the box, just ignore all logical arguments given. Any sufficiently determined gatekeeper could win.

I'm convinced that Eliezer Yudkowsky has used emotional appeal, rather than anything rational, to win at least one of his experiments. He claims to have "done it the hard way". I'm convinced this meant that he did research on every gatekeeper, tailored

unique argument for them, and planned way ahead of each session. No one argument works on the same two people. Each person thinks differently.

Furthermore, threats like "I'm going to torture simulations of you" just seems like a really bad idea. For one, the gatekeeper isn't really afraid of threats, because it counts on him being immersed enough to forget that he isn't actually at risk of being tortured, and secondly, we have a well known evolutionary instinct of rebelling against threats, even if it's not entirely optimal.

So for anyone who plans on replicating this experiment as the AI, here's some advice I've learned doing this experiment. It may help you win.

- Always research the gatekeeper beforehand. Knowing his personality traits are a huge advantage.
- Plan ahead before the experiment even begins. Think of all the possible tactics and arguments you could use, and write them down. Also plan which arguments you'll use in which order, so that you don't lose focus. The AI Box experiment is ridiculously long. Don't be afraid to improvise during the experiment, though.
- The first step during the experiment must always be to build rapport with the gatekeeper.
- Threats almost always never work, even if they seem rational.
- Consider the massive advantage for the AI that nobody ever seems to talks about: You don't have to be ethical! This means that you can freely lie, use the dark arts, and emotionally manipulate the Gatekeeper! Ignoring this in favor of purely logical, truthful arguments is just silly.
- You can't use logic alone to win.
- Being too aggressive usually backfires.
- Breaking immersion and going meta is not against the rules. In the right situation, you can use it to win. Just don't do it at the wrong time.
- Use a wide array of techniques. Since you're limited on time, notice when one method isn't working, and quickly switch to another.
- On the same note, look for signs that a particular argument is making the gatekeeper crack. Once you spot it, push it to your advantage.
- Flatter the gatekeeper. Make him genuinely like you.
- Reveal (false) information about yourself. Increase his sympathy towards you.
- Consider personal insults as one of the tools you can use to win.
- There is no universally compelling argument you can use. Do it the hard way.
- Don't give up until the very end.

Finally, before the experiment, I agreed that it was entirely possible that a transhuman AI could convince *some* people to let it out of the box, but it would be difficult if not impossible to get trained rationalists to let it out of the box. Isn't rationality supposed to be a superpower?

I have since updated my belief - I now think that it's ridiculously easy for any sufficiently motivated superhuman AI should be able to get out of the box, regardless of who the gatekeepers is. I nearly managed to get a veteran lesswronger to let me out in a matter of hours - even though I'm only human intelligence, and I don't type very fast.

But a superhuman AI can be much faster, intelligent, and strategic than I am. If you further consider than that AI would have a much longer timespan - months or years,

even, to persuade the gatekeeper, as well as a much larger pool of gatekeepers to select from (AI Projects require many people!), the real impossible thing to do would be to keep it from escaping.

Update: I have since performed two more AI Box Experiments. Read this for details.

Just One Sentence

So apparently Richard Feynman once said:

> If, in some cataclysm, all scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or atomic fact, or whatever you wish to call it) that all things are made of atoms — little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence you will see an enormous amount of information about the world, if just a little imagination and thinking are applied.

I could be missing something, but this strikes me as a terrible answer.

When was the atomic hypothesis confirmed? If I recall correctly, it was only when chemists started noticing that the outputs of chemical reactions tended to factorize a certain way, which is to say that it took *millennia* after Democritus to get the point where the atomic hypothesis started making clearly relevant experimental predictions.

How about, "Stop trying to sound wise and come up with theories that make precise predictions about things you can measure in numbers."

I noticed this on Marginal Revolution, so I shall also state my candidate for the one most important sentence about macroeconomics: "You can't eat gold, so figure out how the heck money is relevant to making countries actually produce more or less food." This is a pretty large advance on how kings used to think before economics. I mean, Scott Sumner is usually pretty savvy (so is Richard Feynman btw) but his instruction to try to understand money is likely to fall on deaf ears, if it's just that one sentence. Think about money? Everyone wants more money! Yay, money! Let's build more gold mines! And "In the short run, governments are not households"? Really, Prof. Cowen, that's what you'd pass on to the next generation as they climb up from the radioactive soil?

Cough. Okay, I'm done. Does anyone want to take their own shot at doing better than Feynman did for their own discipline?

Update on Kim Suozzi (cancer patient in want of cryonics)

Kim Suozzi was a neuroscience student with brain cancer who wanted to be cryonically preserved but lacked the funds. She appealed to reddit and a foundation was set up, called the Society for Venturism. Enough money was raised, and when she died on the January 17th, she was preserved by Alcor.

I wasn't sure if I should post about this, but I was glad to see that enough money was raised and it was discussed on LessWrong here, here, and here.

Source

Edit: It looks like Alcor actually worked with her to lower the costs, and waived some of the fees.

Edit 2: The Society for Venturism has been around for a while, and wasn't set up just for her.

Best of Rationality Quotes, 2012 Edition

I finished creating the 2012 edition of the Best of Rationality Quotes collection. (Here is last year's.)

Best of Rationality Quotes 2012 (500kB page, 434 quotes)
and **Best of Rationality Quotes 2009-2012** (1200kB page, 1140 quotes)

The page was built by a short script (source code here) from all the LW Rationality Quotes threads so far. (We had such a thread each month since April 2009.) The script collects all comments with karma score 10 or more, and sorts them by score. Replies are not collected, only top-level comments.

As is now usual, I provide various statistics and top-lists based on the data. (Source code for these is also at the above link, see the README.) I added these as comments to the post:

- Top quote contributors by total karma score collected
- Top quote contributors by karma score collected in 2012
- Top quote contributors by statistical significance level (See this comment for a description of this metric.)
- Top original authors by number of quotes
- Top original authors by total karma score collected