# Best of LessWrong: June 2012

1. Ask an experimental physicist
2. The Power of Reinforcement
3. Reply to Holden on 'Tool AI'
4. Nash Equilibria and Schelling Points
5. Plastination is maturing and needs funding, says Hanson
6. Son of Shit Rationalists Say
7. A (small) critique of total utilitarianism
8. Glenn Beck discusses the Singularity, cites SI researchers
9. Intellectual insularity and productivity
10. How to Run a Successful Less Wrong Meetup
11. [Link] Can We Reverse The Stanford Prison Experiment?
12. Introduction to Game Theory: Sequence Guide
13. Conspiracy Theories as Agency Fictions
14. Blogs by LWers
15. [Link] The Greek Heliocentric Theory
16. Bounded versions of Gödel's and Löb's theorems

# Best of LessWrong: June 2012

1. Ask an experimental physicist
2. The Power of Reinforcement
3. Reply to Holden on 'Tool AI'
4. Nash Equilibria and Schelling Points
5. Plastination is maturing and needs funding, says Hanson
6. Son of Shit Rationalists Say
7. A (small) critique of total utilitarianism
8. Glenn Beck discusses the Singularity, cites SI researchers
9. Intellectual insularity and productivity
10. How to Run a Successful Less Wrong Meetup
11. [Link] Can We Reverse The Stanford Prison Experiment?
12. Introduction to Game Theory: Sequence Guide
13. Conspiracy Theories as Agency Fictions
14. Blogs by LWers
15. [Link] The Greek Heliocentric Theory
16. Bounded versions of Gödel's and Löb's theorems

# Ask an experimental physicist

In response to falenas108's "Ask an X" thread. I have a PhD in experimental particle physics; I'm currently working as a postdoc at the University of Cincinnati. Ask me anything, as the saying goes.

This is an experiment. There's nothing I like better than talking about what I do; but I usually find that even quite well-informed people don't know enough to ask questions sufficiently specific that I can answer any better than the next guy. What goes through most people's heads when they hear "particle physics" is, judging by experience, string theory. Well, I dunno nuffin' about string theory - at least not any more than the average layman who has read Brian Greene's book. (Admittedly, neither do string theorists.) I'm equally ignorant about quantum gravity, dark energy, quantum computing, and the Higgs boson - in other words, the big theory stuff that shows up in popular-science articles. For that sort of thing you want a theorist, and not just any theorist at that, but one who works specifically on that problem. On the other hand I'm reasonably well informed about production, decay, and mixing of the charm quark and charmed mesons, but who has heard of that? (Well, now you have.) I know a little about CP violation, a bit about detectors, something about reconstructing and simulating events, a fair amount about how we extract signal from background, and quite a lot about fitting distributions in multiple dimensions.

# The Power of Reinforcement

Part of the sequence: The Science of Winning at Life

Also see: Basics of Animal Reinforcement, Basics of Human Reinforcement, Physical and Mental Behavior, Wanting vs. Liking Revisited, Approving reinforces low-effort behaviors, Applying Behavioral Psychology on Myself.

**Story 1**:

On Skype with Eliezer, I said: "Eliezer, you've been unusually *pleasant* these past three weeks. I'm really happy to see that, and moreover, it increases my probability than an Eliezer-led FAI research team will *work*. What caused this change, do you think?"

Eliezer replied: "Well, three weeks ago I was working with Anna and Alicorn, and every time I said something nice they fed me an M&M."

**Story 2**:

I once witnessed a worker who *hated* keeping a work log because it was only used "against" him. His supervisor would call to say "Why did you spend so much time on *that*?" or "Why isn't *this* done yet?" but never "I saw you handled *X*, great job!" Not surprisingly, he often "forgot" to fill out his worklog.

Ever since I got everyone at the Singularity Institute to keep work logs, I've tried to avoid connections between "concerned" feedback and staff work logs, and instead take time to comment positively on things I see in those work logs.

**Story 3**:

Chatting with Eliezer, I said, "Eliezer, I get the sense that I've inadvertently caused you to be slightly averse to talking to me. Maybe because we disagree on so many things, or something?"

Eliezer's reply was: "No, it's much simpler. Our conversations usually run longer than our previously set deadline, so whenever I finish talking with you I feel drained and slightly cranky."

Now I finish our conversations on time.

**Story 4**:

A major Singularity Institute donor recently said to me: "By the way, I decided that every time I donate to the Singularity Institute, I'll set aside an additional 5% for myself to do fun things with, as a motivation to donate."

### The power of reinforcement

It's amazing to me how consistently we fail to take advantage of the power of reinforcement.

Maybe it's because behaviorist techniques like reinforcement feel like they don't respect human agency enough. But if you aren't treating humans more like animals than *most* people are, then you're *modeling humans poorly.*

You are not an agenty homunculus "corrupted" by heuristics and biases. You just *are* heuristics and biases. And you respond to reinforcement, because most of your motivation systems still work like the motivation systems of other animals.

### A quick reminder of what you learned in high school

- A *reinforcer* is anything that, when it occurs in conjunction with an act, increases the probability that the act will occur again.
- A *positive reinforcer* is something the subject wants, such as food, petting, or praise. *Positive reinforcement* occurs when a target behavior is followed by something the subject wants, and this increases the probability that the behavior will occur again.
- A *negative reinforcer* is something the subject wants to avoid, such as a blow, a frown, or an unpleasant sound. *Negative reinforcement* occurs when a target behavior is followed by some *relief* from something the subject *doesn't want*, and this increases the probability that the behavior will happen again.

### What works

1. **Small reinforcers are fine**, as long as there is a strong correlation between the behavior and the reinforcer (Schneider 1973; Todorov et al. 1984). All else equal, a large reinforcer is more effective than a small one (Christopher 1988; Ludvig et al. 2007; Wolfe 1936), but the more you increase the reinforcer magnitude, the less benefit you get from the increase (Frisch & Dickinson 1990).
2. **The reinforcer should *immediately* follow the target behavior** (Escobar & Bruner 2007; Schlinger & Blakely 1994; Schneider 1990). Pryor (2007) notes that when the reward is food, small bits (like M&Ms) are best because they can be consumed *instantly* instead of being consumed over an extended period of time.
3. **Any *feature* of a behavior can be strengthened** (e.g., its intensity, frequency, rate, duration, persistence, its shape or form), so long as a reinforcer can be made contingent on *that particular feature* (Neuringer 2002).

### Example applications

- If you want someone to call you, then when they *do* call, don't nag them about how they never call you. Instead, be engaging and positive.

- When trying to maintain order in a class, ignore unruly behavior and praise good behavior (Madsen et al. 1968; McNamara 1987).
- Reward originality to encourage creativity (Pryor et al. 1969; Chambers et al. 1977; Eisenberger & Armeli 1997; Eisenberger & Rhoades 2001).
- If you want students to *understand* the material, don't get excited when they guess the teacher's password but instead when they demonstrate a technical understanding.
- To help someone improve at dance or sport, ignore poor performance but reward good performance immediately, for example by shouting "Good!" (Buzas & Allyon 1981) The reason you should ignore poor performance if you say "No, you're doing it wrong!" you are inadvertently punishing the *effort*. A better response to a mistake would be to reinforce the effort: "Good effort! You're almost there! Try once more."
- Reward honesty to help people be more honest with you (Lanza et al 1982).
- Reward opinion-expressing to get people to express their opinions more often (Verplanck 1955).
- You may even be able to reinforce-away annoying *involuntary* behaviors, such as twitches (Laurenti-Lions et al. 1985) or vomiting (Wolf et al. 1965).
- Want a young infant to learn to speak more quickly? Reinforce their attempts at vocalization (Ramely & Finkelstein 1978).
- More training should occur via video games like DragonBox, because computer programs can easily provide *instant* reinforcement *many times a minute* for *very specific behaviors* (Fletcher-Flinn & Gravatt 1995).

For additional examples and studies, see *The Power of Reinforcement* (2004), *Don't Shoot the Dog* (2006), and *Learning and Behavior* (2008).

I close with **Story 5**, from Amy Sutherland:

> For a book I was writing about a school for exotic animal trainers, I started commuting from Maine to California, where I spent my days watching students do the seemingly impossible: teaching hyenas to pirouette on command, cougars to offer their paws for a nail clipping, and baboons to skateboard.
>
> I listened, rapt, as professional trainers explained how they taught dolphins to flip and elephants to paint. Eventually it hit me that the same techniques might work on that stubborn but lovable species, the American husband.
>
> The central lesson I learned from exotic animal trainers is that I should reward behavior I like and ignore behavior I don't. After all, you don't get a sea lion to balance a ball on the end of its nose by nagging. The same goes for the American husband.
>
> Back in Maine, I began thanking Scott if he threw one dirty shirt into the hamper. If he threw in two, I'd kiss him. Meanwhile, I would step over any soiled clothes on the floor without one sharp word, though I did sometimes kick them under the bed. But as he basked in my appreciation, the piles became smaller.
>
> I was using what trainers call "approximations," rewarding the small steps toward learning a whole new behavior...
>
> Once I started thinking this way, I couldn't stop. At the school in California, I'd be scribbling notes on how to walk an emu or have a wolf accept you as a pack

> member, but I'd be thinking, "I can't wait to try this on Scott."
>
> ...After two years of exotic animal training, my marriage is far smoother, my husband much easier to love.

Next post: Rational Romantic Relationships Part 1

Previous post: The Good News of Situationist Psychology

My thanks to Erica Edelman for doing much of the research for this post.

# Reply to Holden on 'Tool AI'

I begin by thanking Holden Karnofsky of Givewell for his rare gift of his detailed, engaged, and helpfully-meant critical article Thoughts on the Singularity Institute (SI). In this reply I will engage with only one of the *many* subjects raised therein, the topic of, as I would term them, non-self-modifying planning Oracles, a.k.a. 'Google Maps AGI' a.k.a. 'tool AI', this being the topic that requires me personally to answer. I hope that my reply will be accepted as addressing the most important central points, though I did not have time to explore every avenue. I certainly do not wish to be logically rude, and if I have failed, please remember with compassion that it's not always obvious to one person what another person will think was the central point.

Luke Mueulhauser and Carl Shulman contributed to this article, but the final edit was my own, likewise any flaws.

## Summary:

Holden's concern is that "SI appears to neglect the potentially important distinction between 'tool' and 'agent' AI." His archetypal example is Google Maps:

> Google Maps is not an *agent*, taking actions in order to maximize a utility parameter. It is a *tool*, generating information and then displaying it in a user-friendly manner for me to consider, use and export or discard as I wish.

The reply breaks down into four heavily interrelated points:

First, Holden seems to think (and Jaan Tallinn doesn't apparently object to, in their exchange) that if a non-self-modifying planning Oracle is indeed the best strategy, then all of SIAI's past and intended future work is wasted. To me it looks like there's a huge amount of overlap in underlying processes in the AI that would have to be built and the insights required to build it, and I would be trying to assemble mostly - though not quite exactly - the same kind of *team* if I was trying to build a non-self-modifying planning Oracle, with the same initial mix of talents and skills.

Second, a non-self-modifying planning Oracle doesn't sound nearly as safe once you stop saying human-English phrases like "describe the consequences of an action to the user" and start trying to come up with math that says scary dangerous things like (he translated into English) "increase the correspondence between the user's belief about relevant consequences and reality". Hence why the people on the team would have to solve the same sorts of problems.

Appreciating the force of the third point is a lot easier if one appreciates the difficulties discussed in points 1 and 2, but is actually empirically verifiable independently: Whether or not a non-self-modifying planning Oracle is the *best* solution in the end, it's not such an *obvious* privileged-point-in-solution-space that someone should be alarmed at SIAI not discussing it. This is empirically verifiable in the sense that 'tool AI' wasn't the obvious solution to e.g. John McCarthy, Marvin Minsky, I. J. Good, Peter Norvig, Vernor Vinge, or for that matter Isaac Asimov. At one point, Holden says:

> One of the things that bothers me most about SI is that there is practically no public content, as far as I can tell, explicitly addressing the idea of a "tool" and

giving arguments for why AGI is likely to work only as an "agent."

If I take literally that this is one of the things that bothers Holden *most...* I think I'd start stacking up some of the literature on the number of different things that *just respectable academics* have suggested as the *obvious solution* to what-to-do-about-AI - none of which would be about non-self-modifying smarter-than-human planning Oracles - and beg him to have some compassion on us for what we *haven't addressed yet*.  It might be the right suggestion, but it's not so obviously right that our failure to prioritize discussing it reflects negligence.

The final point at the end is looking over all the preceding discussion and realizing that, yes, you want to have people specializing in Friendly AI who know this stuff, but as all that preceding discussion is actually the following discussion at this point, I shall reserve it for later.

## 1.  The math of optimization, and the similar parts of a planning Oracle.

What does it take to build a smarter-than-human intelligence, of whatever sort, and have it go well?

A "Friendly AI programmer" is somebody who specializes in seeing the correspondence of mathematical structures to What Happens in the Real World. It's somebody who looks at Hutter's specification of AIXI and reads the actual equations - actually stares at the Greek symbols and not just the accompanying English text - and sees, "Oh, this AI will try to gain control of its reward channel," as well as numerous subtler issues like, "This AI presumes a Cartesian boundary separating itself from the environment; it may drop an anvil on its own head." Similarly, working on TDT means e.g. looking at a mathematical specification of decision theory, and seeing "Oh, this is vulnerable to blackmail" and coming up with a mathematical counter-specification of an AI that isn't so vulnerable to blackmail.

Holden's post seems to imply that if you're building a non-self-modifying planning Oracle (aka 'tool AI') rather than an acting-in-the-world agent, you don't need a Friendly AI programmer because FAI programmers only work on agents. But this isn't how the engineering skills are split up. Inside the AI, whether an agent AI or a planning Oracle, there would be similar AGI-challenges like "build a predictive model of the world", and similar FAI-conjugates of those challenges like finding the 'user' inside an AI-created model of the universe.  The insides would look a lot more similar than the outsides.  An analogy would be supposing that a machine learning professional who does sales optimization for an orange company couldn't possibly do sales optimization for a banana company, because their skills must be about oranges rather than bananas.

Admittedly, if it turns out to be possible to use a human understanding of cognitive algorithms to build and run a smarter-than-human Oracle without it being self-improving - this seems unlikely, but not impossible - then you wouldn't have to solve problems that arise with self-modification.  But this eliminates only one dimension of the work.  And on an even more meta level, it seems like you would call upon almost identical *talents and skills* to come up with whatever insights were required - though if it were predictable in advance that we'd abjure self-modification, then, yes, we'd place less emphasis on e.g. finding a team member with past experience in reflective math, and wouldn't waste (additional) time specializing in reflection.  But if you wanted math inside the planning Oracle that *operated the way you thought it did*, and you wanted somebody who *understood what could possibly go wrong* and how to avoid it, you would need to make a function call to the same sort of talents and skills to build an agent AI, or an Oracle that *was* self-modifying, etc.

## 2. Yes, planning Oracles have hidden gotchas too.

"Tool AI" may sound simple in English, a short sentence in the language of empathically-modeled agents — it's just "a thingy that shows you plans instead of a thingy that goes and does things." If you want to know whether this hypothetical entity does X, you just check whether the outcome of X sounds like "showing someone a plan" or "going and doing things", and you've got your answer. It starts sounding much scarier once you try to say something more formal and internally-causal like "Model the user and the universe, predict the degree of correspondence between the user's model and the universe, and select from among possible explanation-actions on this basis."

Holden, in [his dialogue with Jaan Tallinn](), writes out this attempt at formalizing:

> Here's how I picture the Google Maps AGI ...
>
> ```
> utility_function = construct_utility_function(process_user_input());
>
> foreach $action in $all_possible_actions {
>
>     $action_outcome = prediction_function($action,$data);
>
>     $utility = utility_function($action_outcome);
>
>     if ($utility > $leading_utility) { $leading_utility = $utility;
>
>     $leading_action = $action; }
>
> }
>
> report($leading_action);
> ```
>
> construct_utility_function(process_user_input()) is just a human-quality function for understanding what the speaker wants. prediction_function is an implementation of a human-quality data->prediction function in superior hardware. $data is fixed (it's a dataset larger than any human can process); same with $all_possible_actions. report($leading_action) calls a Google Maps-like interface for understanding the consequences of $leading_action; it basically breaks the action into component parts and displays predictions for different times and conditional on different parameters.

Google Maps doesn't check all possible routes. If I wanted to design Google Maps, I would start out by throwing out a standard planning technique on a connected graph where each edge has a cost function and there's a good heuristic measure of the distance, e.g. [A* search](). If that was too slow, I'd next try some more efficient version like weighted A* (or bidirectional weighted memory-bounded A*, which I expect I could also get off-the-shelf somewhere). Once you introduce weighted A*, you no longer have a guarantee that you're selecting the optimal path. You have a guarantee to within a known factor of the cost of the optimal path — but the actual path selected wouldn't be quite optimal. The suggestion produced would be an approximation whose exact steps depended on the exact algorithm you used. That's true even if you can predict the exact cost — exact utility — of any particular path you actually look at; and even if you have a heuristic that never overestimates the cost.

The reason we don't have God's Algorithm for solving the Rubik's Cube is that there's no perfect way of measuring the distance between any two Rubik's Cube positions — you can't look at two Rubik's cube positions, and figure out the minimum number of moves required to get from one to another. It took 15 years to prove that there was a position requiring at least 20 moves to solve, and then another 15 years to come up with a computer algorithm that could solve any position in at most 20 moves, but we still can't compute the actual, minimum solution to all Cubes ("God's Algorithm"). This, even though we can exactly calculate the cost and consequence of any actual Rubik's-solution-path we consider.

When it comes to AGI — solving general cross-domain "Figure out how to do X" problems — you're not going to get anywhere near the one, true, optimal answer. You're going to — at best, if everything works right — get *a* good answer that's a cross-product of the "utility function" and all the other algorithmic properties that determine what sort of answer the AI finds easy to invent (i.e. can be invented using bounded computing time).

As for the notion that this AGI runs on a "human predictive algorithm" that we got off of neuroscience and then implemented using more computing power, without knowing how it works or being able to enhance it further: It took 30 years of multiple computer scientists doing basic math research, and inventing code, and running that code on a computer cluster, for them to come up with a 20-move solution to the Rubik's Cube. If a planning Oracle is going to produce better solutions than humanity has yet managed to the Rubik's Cube, it needs to be capable of doing original computer science research and writing its own code. You can't get a 20-move solution out of a human brain, using the native human planning algorithm. Humanity can do it, but only by exploiting the ability of humans to explicitly comprehend the deep structure of the domain (not just rely on intuition) and then inventing an artifact, a new design, running code which uses a different and superior cognitive algorithm, to solve that Rubik's Cube in 20 moves. We do all that without being *self*-modifying, but it's still a capability to respect.

And I'm not even going into what it would take for a planning Oracle to out-strategize any human, come up with a plan for persuading someone, solve original scientific problems by looking over experimental data (like Einstein did), design a nanomachine, and so on.

Talking like there's this one simple "predictive algorithm" that we can read out of the brain using neuroscience and overpower to produce better plans... doesn't seem quite congruous with what humanity actually does to produce its predictions and plans.

If we take the concept of the Google Maps AGI at face value, then it actually has four key magical components. (In this case, "magical" isn't to be taken as prejudicial, it's a term of art that means we haven't said how the component works yet.) There's a magical comprehension of the user's utility function, a magical world-model that GMAGI uses to comprehend the consequences of actions, a magical planning element that selects a *non-optimal* path using some method *other* than exploring all possible actions, and a magical explain-to-the-user function.

report($leading_action) isn't exactly a trivial step either. Deep Blue tells you to move your pawn or you'll lose the game. You ask "Why?" and the answer is a gigantic search tree of billions of possible move-sequences, leafing at positions which are heuristically rated using a static-position evaluation algorithm trained on millions of games. Or the planning Oracle tells you that a certain DNA sequence will produce a protein that cures cancer, you ask "Why?", and then humans aren't even capable of verifying, for themselves, the assertion that the peptide sequence will fold into the protein the planning Oracle says it does.

"So," you say, after the first dozen times you ask the Oracle a question and it returns an answer that you'd have to take on faith, "we'll just specify in the utility function that the plan should be understandable."

Whereupon other things start going wrong. Viliam_Bur, in the comments thread, gave this example, which I've slightly simplified:

> Example question: "How should I get rid of my disease most cheaply?" Example answer: "You won't. You will die soon, unavoidably. This report is 99.999% reliable". Predicted human reaction: Decides to kill self and get it over with. Success rate: 100%, the disease is gone. Costs of cure: zero. Mission completed.

Bur is trying to give an example of how things might go wrong if the preference function is over the accuracy of the predictions explained to the human— rather than *just* the human's 'goodness' of the outcome. And if the preference function *was* just over the human's 'goodness' of the end result, rather than the accuracy of the human's understanding of the predictions, the AI might tell you something that was predictively false but whose implementation would lead you to what the AI defines as a 'good' outcome. And if we ask how happy the human is, the resulting decision procedure would exert optimization pressure to convince the human to take drugs, and so on.

I'm not saying any particular failure is 100% certain to occur; rather I'm trying to explain - as handicapped by the need to describe the AI in the native human agent-description language, using empathy to simulate a spirit-in-a-box instead of trying to think in mathematical structures like A* search or Bayesian updating - how, even so, one can still see that the issue is a tad more fraught than it sounds on an immediate examination.

If you see the world just in terms of math, it's even worse; you've got some program with inputs from a USB cable connecting to a webcam, output to a computer monitor, and optimization criteria expressed over some combination of the monitor, the humans looking at the monitor, and the rest of the world. It's a whole lot easier to call what's inside a 'planning Oracle' or some other English phrase than to write a program that does the optimization safely without serious unintended consequences. Show me any attempted specification, and I'll point to the vague parts and ask for clarification in more formal and mathematical terms, and as soon as the design is clarified enough to be a hundred light years from implementation instead of a thousand light years, I'll show a neutral judge how that math would go wrong. (Experience shows that if you try to explain to would-be AGI designers how their design goes wrong, in most cases they just say "Oh, but of course that's not what I meant." Marcus Hutter is a rare exception who specified his AGI in such unambiguous mathematical terms that he actually succeeded at realizing, after some discussion with SIAI personnel, that AIXI would kill off its users and seize control of its reward button. But based on past sad experience with many other would-be designers, I say "Explain to a neutral judge how the math kills" and not "Explain to the person who invented that math and likes it.")

Just as the gigantic gap between smart-sounding English instructions and actually smart algorithms is the main source of difficulty in AI, there's a gap between benevolent-sounding English and actually benevolent algorithms which is the source of difficulty in FAI. "Just make suggestions - don't *do* anything!" is, in the end, just more English.

## 3. Why we haven't already discussed Holden's suggestion

> One of the things that bothers me most about SI is that there is practically no public content, as far as I can tell, explicitly addressing the idea of a "tool" and giving arguments for why AGI is likely to work only as an "agent."

The above statement seems to lack perspective on how *many* different things various people see as *the one obvious solution* to Friendly AI. Tool AI wasn't the obvious solution to John McCarthy, I.J. Good, or Marvin Minsky. Today's leading AI textbook, *Artificial Intelligence: A Modern Approach* - where you can learn all about A* search, by the way - discusses Friendly AI and AI risk for 3.5 pages but doesn't mention tool AI as an obvious solution. For Ray Kurzweil, the obvious solution is merging humans and AIs. For Jurgen Schmidhuber, the obvious solution is AIs that value a certain complicated definition of complexity in their sensory inputs. Ben Goertzel, J. Storrs Hall, and Bill Hibbard, among others, have all written about how silly Singinst is to pursue Friendly AI when the solution is obviously X, for various different X. Among current leading people working on serious AGI programs labeled as such, neither Demis Hassabis (VC-funded to the tune of several million dollars) nor Moshe Looks (head of AGI research at Google) nor Henry Markram (Blue Brain at IBM) think that the obvious answer is Tool AI. Vernor Vinge, Isaac Asimov, and any number of other SF writers with technical backgrounds who spent serious time thinking about these issues didn't converge on that solution.

Obviously I'm not saying that nobody should be allowed to propose solutions because someone else would propose a different solution. I have been known to advocate for particular developmental pathways for Friendly AI myself. But I haven't, for example, told Peter Norvig that deterministic self-modification is such an obvious solution to Friendly AI that I would mistrust his whole AI textbook if he didn't spend time discussing it.

At one point in his conversation with Tallinn, Holden argues that AI will inevitably be developed along planning-Oracle lines, because making suggestions to humans is the natural course that most software takes. Searching for counterexamples instead of positive examples makes it clear that most lines of code don't do this. Your computer, when it reallocates RAM, doesn't pop up a button asking you if it's okay to reallocate RAM in such-and-such a fashion. Your car doesn't pop up a suggestion when it wants to change the fuel mix or apply dynamic stability control. Factory robots don't operate as human-worn bracelets whose blinking lights suggest motion. High-frequency trading programs execute stock orders on a microsecond timescale. Software that does happen to interface with humans is selectively visible and salient to humans, especially the tiny part of the software that does the interfacing; but this is a special case of a general cost/benefit tradeoff which, more often than not, turns out to swing the other way, because human advice is either too costly or doesn't provide enough benefit. Modern AI programmers are generally more interested in e.g. pushing the technological envelope to allow self-driving cars than to "just" do Google Maps. Branches of AI that invoke human aid, like hybrid chess-playing algorithms designed to incorporate human advice, are a field of study; but they're the exception rather than the rule, and occur primarily where AIs can't yet do something humans do, e.g. humans acting as oracles for theorem-provers, where the humans suggest a route to a proof and the AI actually follows that route. This is another reason why planning Oracles were not a uniquely obvious solution to the various academic AI researchers, would-be AI-creators, SF writers, etcetera, listed above. Again, regardless of whether a planning Oracle is actually the best solution, Holden seems to be empirically-demonstrably overestimating the degree to which other people will automatically have his preferred solution come up first in their search ordering.

## 4. Why we should have full-time Friendly AI specialists just like we have trained professionals doing anything else mathy that somebody actually cares about getting right, like pricing interest-rate options or something

I hope that the preceding discussion has made, by example instead of mere argument, what's probably the most important point: If you want to have a sensible discussion about which AI designs are safer, there are specialized skills you can apply to that discussion, as built up over years of study and practice by someone who specializes in answering that sort of question.

This isn't meant as an argument from authority. It's not meant as an attempt to say that only experts should be allowed to contribute to the conversation. But it is meant to say that there is (and ought to be) room in the world for Friendly AI specialists, just like there's room in the world for specialists on optimal philanthropy (e.g. Holden).

The decision to build a non-self-modifying planning Oracle would be properly made by someone who: understood the risk gradient for self-modifying vs. non-self-modifying programs; understood the risk gradient for having the AI thinking about the thought processes of the human watcher and trying to come up with plans implementable by the human watcher in the service of locally absorbed utility functions, vs. trying to implement its own plans in the service of more globally descriptive utility functions; and who, above all, understood on a technical level what exactly gets *accomplished* by having the plans routed through a human. I've given substantial previous thought to describing more precisely what happens — what is being gained, and how much is being gained — when a human "approves a suggestion" made by an AI. But that would be another a different topic, plus I haven't made too much progress on saying it precisely anyway.

In the transcript of Holden's conversation with Jaan Tallinn, it looked like Tallinn didn't deny the assertion that Friendly AI skills would be inapplicable if we're building a Google Maps AGI. I would deny that assertion and emphasize that denial, because to me it seems that it is exactly Friendly AI programmers who would be able to tell you if the risk gradient for non-self-modification vs. self-modification, the risk gradient for routing plans through humans vs. acting as an agent, the risk gradient for requiring human approval vs. unapproved action, and the actual feasibility of directly constructing transhuman modeling-prediction-and-planning algorithms through directly design of sheerly better computations than are presently run by the human brain, had the right combination of properties to imply that you ought to go construct a non-self-modifying planning Oracle. Similarly if you wanted an AI that took a limited set of actions in the world with human approval, or if you wanted an AI that "just answered questions instead of making plans".

It is similarly implied that a "philosophical AI" might obsolete Friendly AI programmers. If we're talking about PAI that can start with a human's terrible decision theory and come up with a good decision theory, or PAI that can start from a human talking about bad metaethics and then construct a good metaethics... I don't want to say "impossible", because, after all, that's just what human philosophers do. But we are not talking about a trivial invention here. Constructing a "philosophical AI" is a Holy Grail precisely because it's FAI-complete (just ask it "What AI should we build?"), and has been discussed (e.g. with and by Wei Dai) over the years on the old SL4 mailing list and the modern Less Wrong. But it's really not at all clear how you could write an algorithm which would knowably produce the correct answer to the entire puzzle of anthropic reasoning, without being in possession of that correct answer yourself (in the same way that we can have Deep Blue win chess games without knowing the exact moves, but understanding exactly what abstract work Deep Blue is doing to solve the problem).

Holden's post presents a restrictive view of what "Friendly AI" people are supposed to learn and know — that it's about machine learning for optimizing orange sales but not apple sales, or about producing an "agent" that implements CEV — which is something of a straw view, much weaker than the view that a Friendly AI programmer takes of Friendly AI programming. What the human species needs from an x-risk perspective is experts on This Whole Damn Problem, who will acquire whatever skills are needed to that end. The Singularity Institute exists to host such people and enable their research—once we have enough funding to find and recruit them.  See also, How to Purchase AI Risk Reduction.

I'm pretty sure Holden has met people who think that having a whole institute to rate the efficiency of charities is pointless overhead, especially people who think that their own charity-solution is too obviously good to have to contend with busybodies pretending to specialize in thinking about 'marginal utility'.  Which Holden knows about, I would guess, from being paid quite well to think about that economic details when he was a hedge fundie, and learning from books written by professional researchers before then; and the really key point is that people who haven't studied all that stuff don't even realize what they're missing by trying to wing it.  If you don't know, you don't know *what* you don't know, or the cost of not knowing.  Is there a problem of figuring out who might know something you don't, if Holden insists that there's this strange new stuff called 'marginal utility' you ought to learn about?  Yes, there is.  But is someone who trusts their philanthropic dollars to be steered just by the warm fuzzies of their heart, doing something wrong?  Yes, they are.  It's one thing to say that SIAI isn't known-to-you to be doing it right - another thing still to say that SIAI is known-to-you to be doing it wrong - and then quite another thing entirely to say that there's no need for Friendly AI programmers *and you know it,* that anyone can see it without resorting to math or cracking a copy of AI: A Modern Approach.  I do wish that Holden would at least credit that the task SIAI is taking on contains at least as many gotchas, relative to the instinctive approach, as optimal philanthropy compared to instinctive philanthropy, and might likewise benefit from some full-time professionally specialized attention, just as our society creates trained professionals to handle any other problem that someone actually cares about getting right.

On the other side of things, Holden says that *even if* Friendly AI is proven and checked:

> "I believe that the probability of an unfavorable outcome - by which I mean an outcome essentially equivalent to what a UFAI would bring about - exceeds 90% in such a scenario."

It's nice that this appreciates that the problem is hard.  Associating all of the difficulty with agenty proposals and thinking that it goes away as soon as you invoke tooliness is, well, of this I've already spoken. I'm not sure whether this irreducible-90%-doom assessment is based on a common straw version of FAI where all the work of the FAI programmer goes into "proving" something and doing this carefully checked proof which then - alas, poor Spock! - turns out to be no more relevant than proving that the underlying CPU does floating-point arithmetic correctly if the transistors work as stated. I've repeatedly said that the idea behind proving determinism of self-modification isn't that this guarantees safety, but that if you prove the self-modification stable the AI *might* work, whereas if you try to get by with no proofs at all, doom is *guaranteed*. My mind keeps turning up Ben Goertzel as the one who invented this caricature - "Don't you understand, poor fool Eliezer, life is full of uncertainty, your attempt to flee from it by refuge in 'mathematical proof' is doomed" - but I'm not sure he was actually the inventor. In any case, the burden of safety isn't carried just by the proof, it's carried mostly by proving the right thing. If Holden is assuming that we're just running away from the inherent uncertainty of life by taking refuge in mathematical proof, then, yes, 90% probability of doom is an understatement, the vast majority of plausible-on-first-glance goal criteria you can prove stable will also kill you.

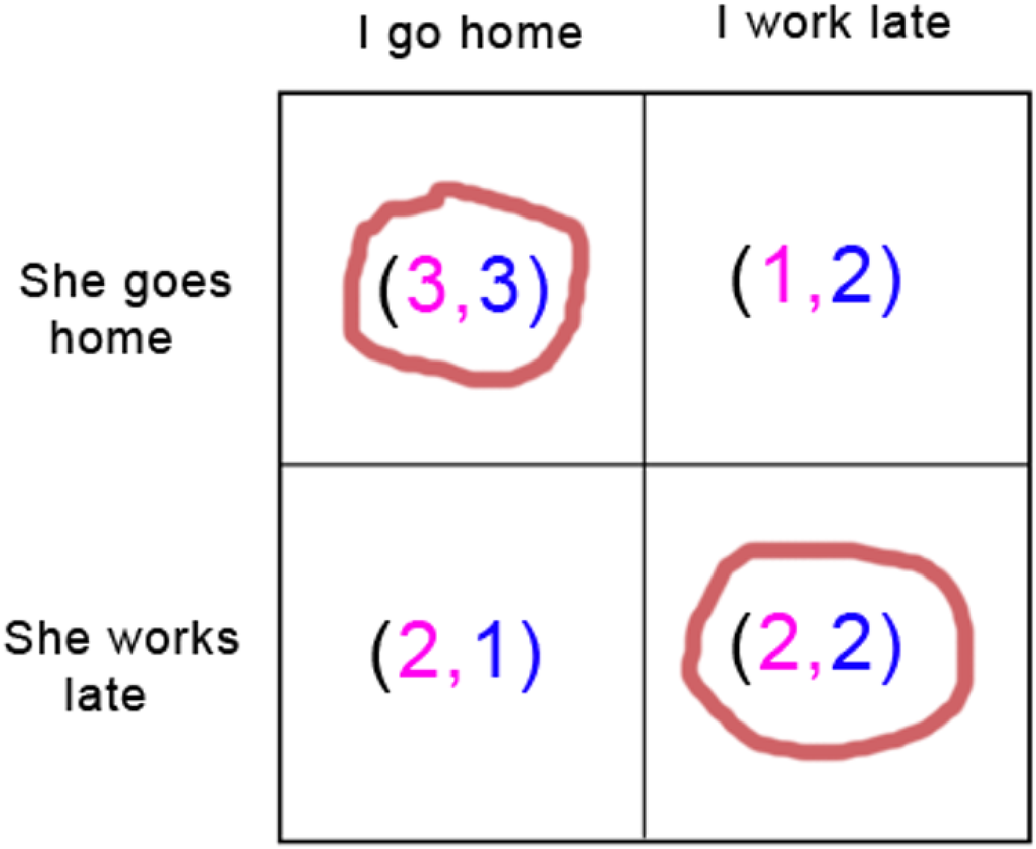
If Holden's assessment does take into account a great effort to select the right theorem to prove - and attempts to incorporate the difficult but finitely difficult feature of meta-level error-detection, as it appears in e.g. the CEV proposal - and he is still assessing 90% doom probability, then I must ask, "What do you think you know and how do you think you know it?" The complexity of the human mind is finite; there's only so many things we want or would-want. Why would someone claim to know that proving the right thing is beyond human ability, even if "100 of the world's most intelligent and relevantly experienced people" (Holden's terms) check it over? There's hidden complexity of wishes, but not infinite complexity of wishes or unlearnable complexity of wishes. There are deep and subtle gotchas but not an unending number of them. And if that *were* the setting of the hidden variables - how would you end up knowing that with 90% probability in advance? I don't mean to wield my own ignorance as a sword or engage in motivated uncertainty - I hate it when people argue that if they don't know something, nobody else is allowed to know either - so please note that I'm also counterarguing from positive facts pointing the other way: the human brain is complicated but not infinitely complicated, there are hundreds or thousands of cytoarchitecturally distinct brain areas but not trillions or googols.  If humanity had two hundred years to solve FAI using human-level intelligence *and there was no penalty for guessing wrong* I would be pretty relaxed about the outcome.  If Holden says there's 90% doom probability left over no matter what sane intelligent people do (all of which goes away if you just build Google Maps AGI, but leave that aside for now) I would ask him what he knows now, in advance, that all those sane intelligent people will miss.  I don't see how you could (well-justifiedly) access that epistemic state.

I acknowledge that there are points in Holden's post which are not addressed in this reply, acknowledge that these points are also deserving of reply, and hope that other SIAI personnel will be able to reply to them.

# Nash Equilibria and Schelling Points

A Nash equilibrium is an outcome in which neither player is willing to unilaterally change her strategy, and they are often applied to games in which both players move simultaneously and where decision trees are less useful.

Suppose my girlfriend and I have both lost our cell phones and cannot contact each other. Both of us would really like to spend more time at home with each other (utility 3). But both of us also have a slight preference in favor of working late and earning some overtime (utility 2). If I go home and my girlfriend's there and I can spend time with her, great. If I stay at work and make some money, that would be pretty okay too. But if I go home and my girlfriend's not there and I have to sit around alone all night, that would be the worst possible outcome (utility 1). Meanwhile, my girlfriend has the same set of preferences: she wants to spend time with me, she'd be okay with working late, but she doesn't want to sit at home alone.

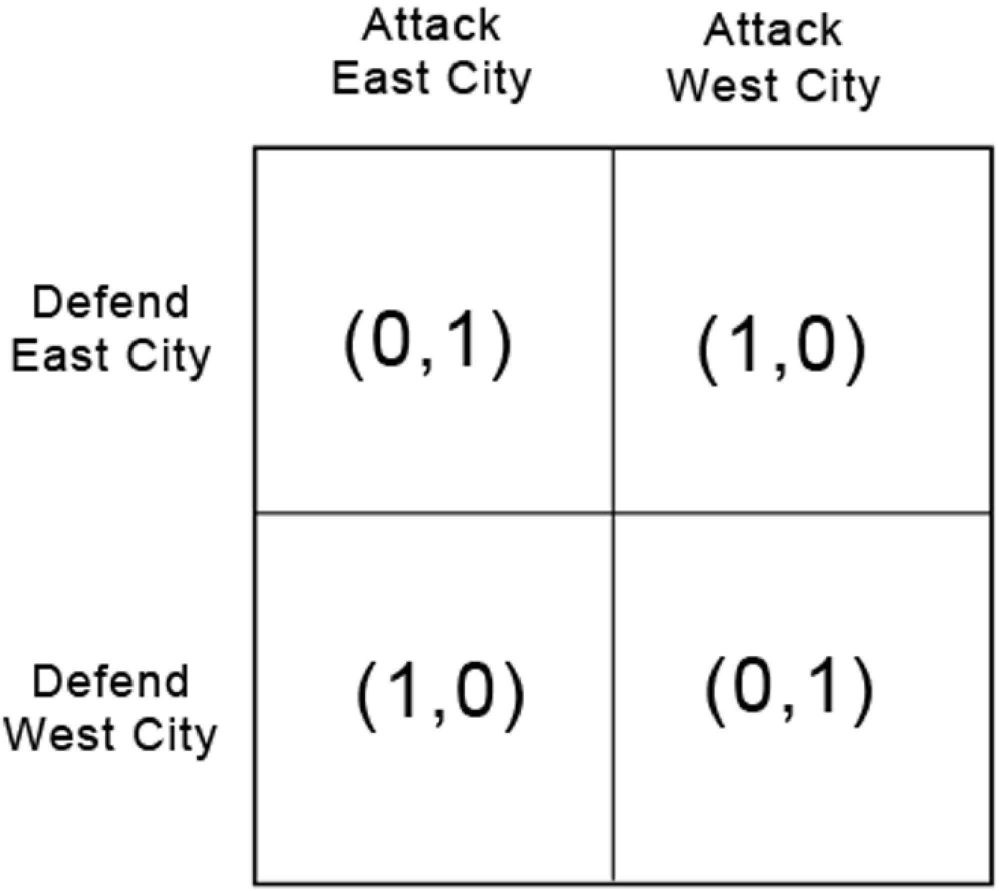


| | I go home | I work late |
|---|---|---|
| She goes home | (3,3) | (1,2) |
| She works late | (2,1) | (2,2) |

This “game” has two Nash equilibria. If we both go home, neither of us regrets it: we can spend time with each other and we've both got our highest utility. If we both stay at work, again, neither of us regrets it: since my girlfriend is at work, I am glad I stayed at work instead of going home, and since I am at work, my girlfriend is glad she stayed at work instead of going home. Although we both may wish that we had both gone home, neither of us specifically regrets our own choice, given our knowledge of how the other acted.

When all players in a game are reasonable, the (apparently) rational choice will be to go for a Nash equilibrium (why would you want to make a choice you'll regret when you know what the other player chose?) And since John Nash (remember that movie *A Beautiful Mind*?) proved that every game has at least one, all games between well-informed rationalists (who are not also being superrational in a sense to be discussed later) should end in one of these.

What if the game seems specifically designed to thwart Nash equilibria? Suppose you are a general invading an enemy country's heartland. You can attack one of two targets, East City or West City (you declared war on them because you were offended by their uncreative toponyms). The enemy general only has enough troops to defend one of the two cities. If you attack an undefended city, you can capture it easily, but if you attack the city with the enemy army, they will successfully fight you off.

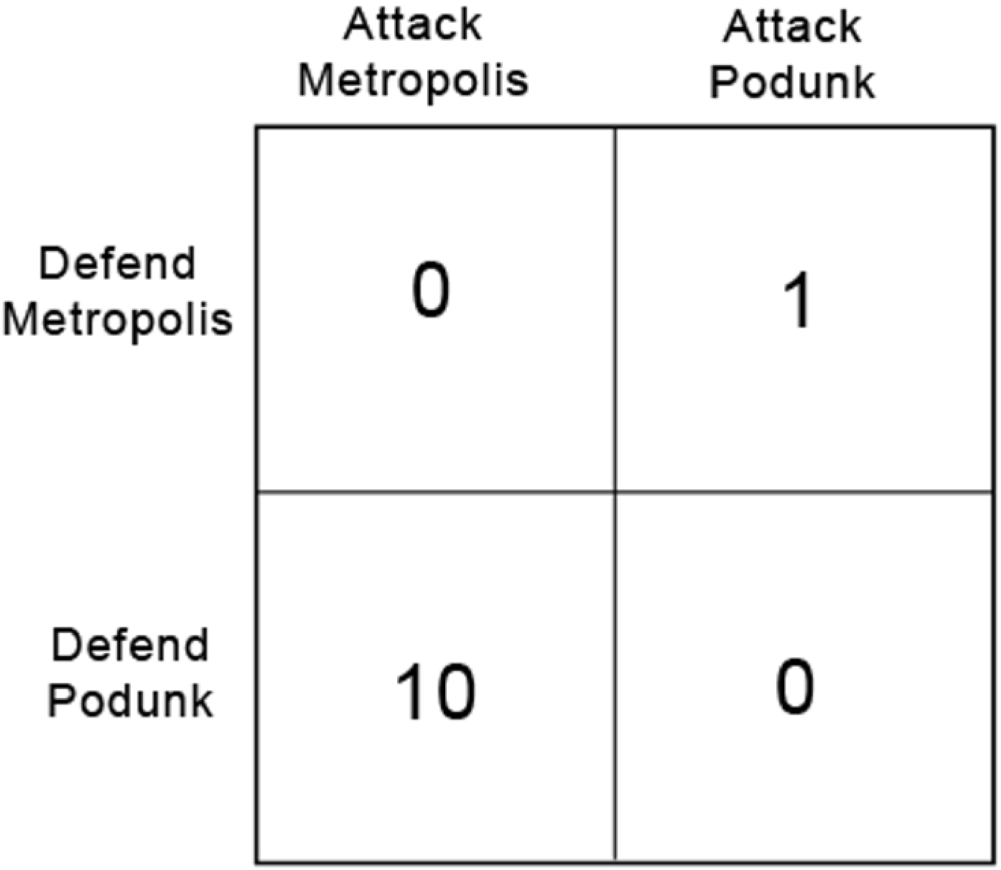


|  | Attack East City | Attack West City |
|---|---|---|
| Defend East City | (0,1) | (1,0) |
| Defend West City | (1,0) | (0,1) |

Here there is no Nash equilibrium without introducing randomness. If both you and your enemy choose to go to East City, you will regret your choice - you should have gone to West and taken it undefended. If you go to East and he goes to West, he will regret his choice - he should have gone East and stopped you in your tracks. Reverse the names, and the same is true of the branches where you go to West City. So every option has someone regretting their choice, and there is no simple Nash equilibrium. What do you do?

Here the answer should be obvious: it doesn't matter. Flip a coin. If you flip a coin, and your opponent flips a coin, neither of you will regret your choice. Here we see a "mixed Nash equilibrium", an equilibrium reached with the help of randomness.

We can formalize this further. Suppose you are attacking a different country with two new potential targets: Metropolis and Podunk. Metropolis is a rich and strategically important city (utility: 10); Podunk is an out of the way hamlet barely worth the trouble of capturing it (utility: 1).



| | Attack Metropolis | Attack Podunk |
|---|---|---|
| Defend Metropolis | 0 | 1 |
| Defend Podunk | 10 | 0 |

A so-called first-level player thinks: “Well, Metropolis is a better prize, so I might as well attack that one. That way, if I win I get 10 utility instead of 1”

A second-level player thinks: “Obviously Metropolis is a better prize, so my enemy expects me to attack that one. So if I attack Podunk, he'll never see it coming and I can take the city undefended.”

A third-level player thinks: “Obviously Metropolis is a better prize, so anyone clever would never do something as obvious as attack there. They'd attack Podunk instead. But my opponent knows that, so, seeking to stay one step ahead of me, he has defended Podunk. He will never expect me to attack Metropolis, because that would be too obvious. Therefore, the city will actually be undefended, so I should take Metropolis.”

And so on ad infinitum, until you become hopelessly confused and have no choice but to spend years developing a resistance to iocane powder.

But surprisingly, there is a single best solution to this problem, even if you are playing against an opponent who, like Professor Quirrell, plays “one level higher than you.”

When the two cities were equally valuable, we solved our problem by flipping a coin. That won't be the best choice this time. Suppose we flipped a coin and attacked Metropolis when we got heads, and Podunk when we got tails. Since my opponent can predict my strategy, he would defend Metropolis every time; I am equally likely to attack Podunk and Metropolis, but taking Metropolis would cost them much more utility. My total expected utility from flipping the coin is 0.5: half the time I successfully take Podunk and gain 1 utility, and half the time I am defeated at Metropolis and gain 0.And this is not a Nash equilibrium: if I had known my opponent's strategy was to defend Metropolis every time, I would have skipped the coin flip and gone straight for Podunk.

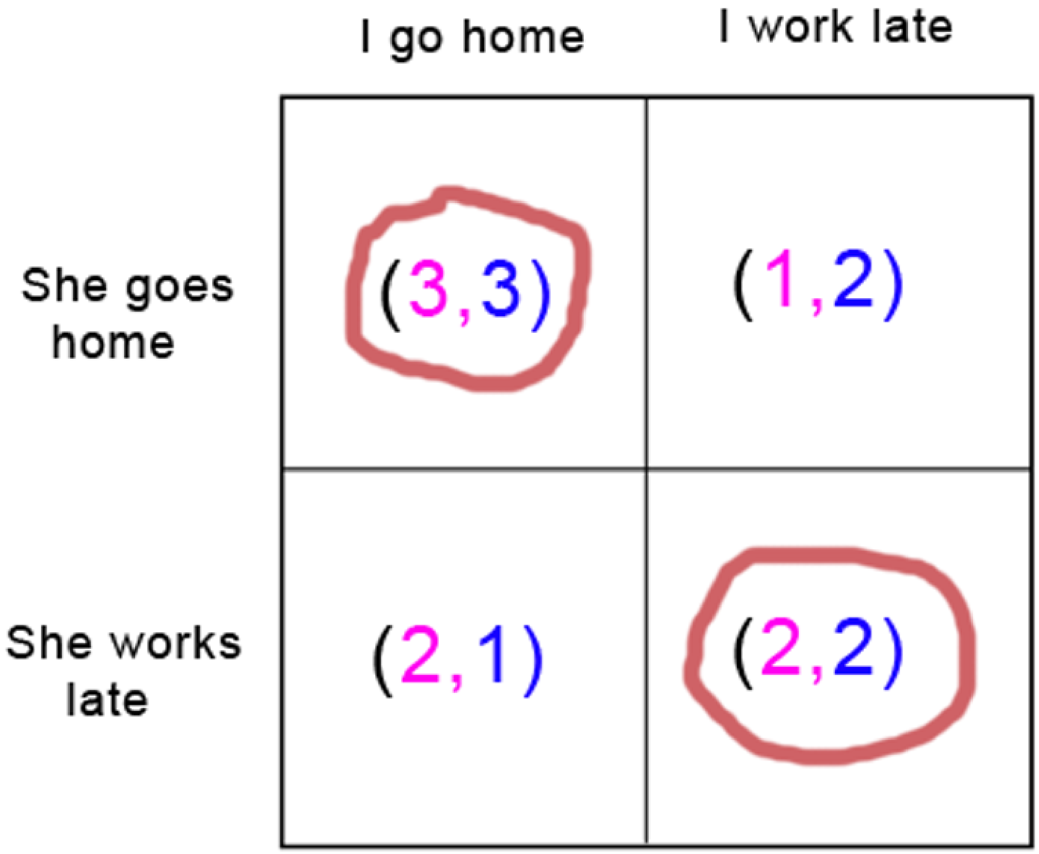
So how can I find a Nash equilibrium? In a Nash equilibrium, I don't regret my strategy when I learn my opponent's action. If I can come up with a strategy that pays exactly the same utility whether my opponent defends Podunk or Metropolis, it will have this useful property. We'll start by supposing I am flipping a *biased* coin that lands on Metropolis x percent of the time, and therefore on Podunk (1-x) percent of the time. To be truly indifferent which city my opponent defends, 10x (the utility my strategy earns when my opponent leaves Metropolis undefended) should equal 1(1-x) (the utility my strategy earns when my opponent leaves Podunk undefended). Some quick algebra finds that 10x = 1(1-x) is satisfied by x = 1/11. So I should attack Metropolis 1/11 of the time and Podunk 10/11 of the time.

My opponent, going through a similar process, comes up with the suspiciously similar result that he should defend Metropolis 10/11 of the time, and Podunk 1/11 of the time.

If we both pursue our chosen strategies, I gain an average 0.9090... utility each round, soundly beating my previous record of 0.5, and my opponent suspiciously loses an average -.9090 utility. It turns out there is no other strategy I can use to consistently do better than this when my opponent is playing optimally, and that even if I knew my opponent's strategy I would not be able to come up with a better strategy to beat it. It also turns out that there is no other strategy my opponent can use to consistently do better than this if I am playing optimally, and that my opponent, upon learning my strategy, doesn't regret his strategy either.

In *The Art of Strategy*, Dixit and Nalebuff cite a real-life application of the same principle in, of all things, penalty kicks in soccer. A right-footed kicker has a better chance of success if he kicks to the right, but a smart goalie can predict that and will defend to the right; a player expecting this can accept a less spectacular kick to the left if he thinks the left will be undefended, but a very smart goalie can predict this too, and so on. Economist Ignacio Palacios-Huerta laboriously analyzed the success rates of various kickers and goalies on the field, and found that they actually pursued a mixed strategy generally within 2% of the game theoretic ideal, proving that people are pretty good at doing these kinds of calculations unconsciously.

So every game really does have at least one Nash equilibrium, even if it's only a mixed strategy. But some games can have many, many more. Recall the situation between me and my girlfriend:



|  | I go home | I work late |
|---|---|---|
| She goes home | (3,3) | (1,2) |
| She works late | (2,1) | (2,2) |

There are two Nash equilibria: both of us working late, and both of us going home. If there were only one equilibrium, and we were both confident in each other's rationality, we could choose that one and there would be no further problem. But in fact this game does present a problem: intuitively it seems like we might still make a mistake and end up in different places.

Here we might be tempted to just leave it to chance; after all, there's a 50% probability we'll both end up choosing the same activity. But other games might have thousands or millions of possible equilibria and so will require a more refined approach.

*Art of Strategy* describes a game show in which two strangers were separately taken to random places in New York and promised a prize if they could successfully meet up; they had no communication with one another and no clues about how such a meeting was to take place. Here there are a nearly infinite number of possible choices: they could both meet at the corner of First Street and First Avenue at 1 PM, they could both meet at First Street and Second Avenue at 1:05 PM, etc. Since neither party would regret their actions (if I went to First and First at 1 and found you there, I would be thrilled) these are all Nash equilibria.

Despite this mind-boggling array of possibilities, in fact all six episodes of this particular game ended with the two contestants meeting successfully after only a few days. The most popular meeting site was the Empire State Building at noon.

How did they do it? The world-famous Empire State Building is what game theorists call focal: it stands out as a natural and obvious target for coordination. Likewise noon, classically considered the very middle of the day, is a focal point in time. These focal points, also called

Schelling points after theorist Thomas Schelling who discovered them, provide an obvious target for coordination attempts.

What makes a Schelling point? The most important factor is that it be special. The Empire State Building, depending on when the show took place, may have been the tallest building in New York; noon is the only time that fits the criteria of “exactly in the middle of the day”, except maybe midnight when people would be expected to be too sleepy to meet up properly.

Of course, specialness, like beauty, is in the eye of the beholder. David Friedman writes:

> *Two people are separately confronted with the list of numbers [2, 5, 9, 25, 69, 73, 82, 96, 100, 126, 150 ] and offered a reward if they independently choose the same number. If the two are mathematicians, it is likely that they will both choose 2—the only even prime. Non-mathematicians are likely to choose 100—a number which seems, to the mathematicians, no more unique than the other two exact squares. Illiterates might agree on 69, because of its peculiar symmetry—as would, for a different reason, those whose interest in numbers is more prurient than mathematical.*

A recent open thread comment pointed out that you can justify anything with “for decision-theoretic reasons” or “due to meta-level concerns”. I humbly propose adding “as a Schelling point” to this list, except that the list is tongue-in-cheek and Schelling points really do explain almost everything - stock markets, national borders, marriages, private property, religions, fashion, political parties, peace treaties, social networks, software platforms and languages all involve or are based upon Schelling points. In fact, whenever something has “symbolic value” a Schelling point is likely to be involved in some way. I hope to expand on this point a bit more later.

Sequential games can include one more method of choosing between Nash equilibria: the idea of a subgame-perfect equilibrium, a special kind of Nash equlibrium that remains a Nash equilibrium for every subgame of the original game. In more intuitive terms, this equilibrium means that even in a long multiple-move game no one at any point makes a decision that goes against their best interests (remember the example from the last post, where we crossed out the branches in which Clinton made implausible choices that failed to maximize his utility?) Some games have multiple Nash equilibria but only one subgame-perfect one; we'll examine this idea further when we get to the iterated prisoners' dilemma and ultimatum game.

In conclusion, every game has at least one Nash equilibrium, a point at which neither player regrets her strategy even when she knows the other player's strategy. Some equilibria are simple choices, others involve plans to make choices randomly according to certain criteria. Purely rational players will always end up at a Nash equilibrium, but many games will have multiple possible equilibria. If players are trying to coordinate, they may land at a Schelling point, an equilibria which stands out as special in some way.

# Plastination is maturing and needs funding, says Hanson

This is a linkpost for http://www.overcomingbias.com/2012/06/plastination-is-near.html

> Though cryonics has been practiced for forty years, its techniques have improved only slowly; its few customers can only induce a tiny research effort. The much larger brain research community, in contrast, has been rapidly improving their ways to do fast cheap detailed 3D brain scans, and to prepare samples for such scans. You see, brain researchers need ways to stop brain samples from changing, and to be strong against scanning disruptions, just so they can study brain samples at their leisure.
>
> These brain research techniques have now reached two key milestones:
>
> 1. They've found new ways to "fix" brain samples by filling them with plastic, ways that seem impressively reliable, resilient, and long lasting, and which work on large brain volumes (e.g., here). Such plastination techniques seem close to being able to save enough info in entire brains for centuries, without needing continual care. Just dumping a plastic brain in a box in a closet might work fine.
> 2. Today, for a few tens of thousands of dollars, less than the price charged for one cryonics customer, it is feasible to have independent lab(s) take random samples from whole mouse or human brains preserved via either cryonics or plastination, and do high (5nm) resolution 3D scans to map out thousands of neighboring cells, their connections, and connection strengths, to test if either of these approaches clearly preserve such key brain info.
>
> An anonymous donor has actually funded a $100K Brain Preservation Prize, paid to the first team(s) to pass this test on a human brain, with a quarter of the prize going to those that first pass the test on a mouse brain. Cryonics and plastination teams have already submitted whole mouse brains to be tested. The only hitch is that the prize organization needs money (~25-50K$) to actually do the tests!

Comments?  If superior brain preservation can be demonstrated under a 5nm-resolution 3D scan, plastination wins over vitrification hands-down.  Is Robin missing anything here, or is this indeed as important as he says?

# Son of Shit Rationalists Say

A long time ago, in the colder seasons, I asked for suggestions for a Shit Rationalists Say video. Due to other concerns it took me this long to put it together, and the meme has long since passed. However, here it is.

Shit Rationalists Say

It is my first time in front of a camera, so I'm shakey. But I learned, and there it is.

# A (small) critique of total utilitarianism

In total utilitarianism, it is a morally neutral act to kill someone (in a painless and unexpected manner) and creating/giving birth to another being of comparable happiness (or preference satisfaction or welfare). In fact if one can kill a billion people to create a billion and one, one is morally compelled to do so. And this is true for real people, not just thought experiment people - living people with dreams, aspirations, grudges and annoying or endearing quirks. To avoid causing extra pain to those left behind, it is better that you kill off whole families and communities, so that no one is left to mourn the dead. In fact the most morally compelling act would be to kill off the whole of the human species, and replace it with a slightly larger population.

We have many real world analogues to this thought experiment. For instance, it seems that there is only a small difference between the happiness of richer nations and poorer nations, while the first consume many more resources than the second. Hence to increase utility we should simply kill off all the rich, and let the poor multiply to take their place (continually bumping off any of the poor that gets too rich). Of course, the rich world also produces most of the farming surplus and the technology innovation, which allow us to support a larger population. So we should aim to kill everyone in the rich world apart from farmers and scientists - and enough support staff to keep these professions running (Carl Shulman correctly points out that we may require most of the rest of the economy as "support staff". Still, it's very likely that we could kill off a significant segment of the population - those with the highest consumption relative to their impact of farming and science - and still "improve" the situation).

Even if turns out to be problematic to implement in practice, a true total utilitarian should be thinking: "I really, really wish there was a way to do targeted killing of many people in the USA, Europe and Japan, large parts of Asia and Latin America and some parts of Africa - it makes me sick to the stomach to think that I can't do that!" Or maybe: "I really really wish I could make everyone much poorer without affecting the size of the economy - I wake up at night with nightmare because these people remain above the poverty line!"

I won't belabour the point. I find those actions personally repellent, and I believe that nearly everyone finds them somewhat repellent or at least did so at some point in their past. This doesn't mean that it's the wrong thing to do - after all, the accepted answer to the torture vs dust speck dilemma feels intuitively wrong, at least the first time. It does mean, however, that there must be very strong countervailing arguments to balance out this initial repulsion (maybe even a mathematical theorem). For without that... how to justify all this killing?

Hence for the rest of this post, I'll be arguing that total utilitarianism is built on a foundation of dust, and thus provides no reason to go against your initial intuitive judgement in these problems. The points will be:

1. Bayesianism and the fact that you should follow a *utility function* in no way compel you towards total *utilitarianism*. The similarity in names does not mean the concepts are on similarly rigorous foundations.
2. Total utilitarianism is neither a simple, nor an elegant theory. In fact, it is under-defined and arbitrary.

3. The most compelling argument for total utilitarianism (basically the one that establishes the repugnant conclusion), is a very long chain of imperfect reasoning, so there is no reason for the conclusion to be solid.
4. Considering the preferences of non-existent beings does not establish total utilitarianism.
5. When considering competing moral theories, total utilitarianism does not "win by default" thanks to its large values as the population increases.
6. Population ethics is *hard*, just as normal ethics is.

## A utility function does not compel total (or average) utilitarianism

There are strong reasons to suspect that the best decision process is one that maximises expected utility for a particular utility function. Any process that does not do so, leaves itself open to be money pumped or taken advantage of. This point has been reiterated again and again on Less Wrong, and rightly so.

Your utility function must be over states of the universe - but that's the only restriction. The theorem says nothing further about the content of your utility function. If you prefer a world with a trillion ecstatic super-humans to one with a septillion subsistence farmers - or vice versa - then as long you maximise your expected utility, the money pumps can't touch you, and the standard Bayesian arguments don't influence you to change your mind. Your values are fully rigorous.

For instance, in the torture vs dust speck scenario, average utilitarianism also compels you to take torture, as do a host of other possible utility functions. A lot of arguments around this subject, that may implicitly feel to be in favour of total utilitarianism, turn out to be nothing of the sort. For instance, avoiding scope insensitivity does not compel you to total utilitarianism, and you can perfectly allow birth-death asymmetries or similar intuitions, while remaining an expected utility maximiser.

## Total utilitarianism is not simple nor elegant, but is arbitrary

Total utilitarianism is defined as maximising the sum of everyone's individual utility function. That's a simple definition. But what are these individual utility functions? Do people act like expected utility maximisers? In a word... no. In another word... NO. In yet another word... NO!

So what are these utilities? Are they the utility that the individuals "should have"? According to what and who's criteria? Is it "welfare"? How is that defined? Is it happiness? Again, how is that defined? Is it preferences? On what scale? And what if the individual disagrees with the utility they are supposed to have? What if their revealed preferences are different again?

There are (various different) ways to start resolving these problems, and philosophers have spent a lot of ink and time doing so. The point remains that total utilitarianism cannot claim to be a simple theory, if the objects that it sums over are so poorly and controversially defined.

And the sum itself is a huge problem. There is no natural scale on which to compare utility functions. Divide one utility function by a billion, multiply the other by $e^{\pi}$, and they are still perfectly valid utility functions. In a study group at the FHI, we've been looking at various ways of combining utility functions - equivalently, of doing interpersonal utility comparisons (IUC). Turns out it's very hard, there seems no natural way of doing this, and a lot has also been written about this, concluding little. Unless your theory comes with a particular IUC method, the only way of summing these utilities is to do an essentially arbitrary choice for each individual before summing. Thus standard total utilitarianism is an arbitrary sum of ill defined, non-natural objects.

Why then is it so popular? Well, one reason is that there are models that make use of something like total utilitarianism to great effect. Classical economic theory, for instance, models everyone as perfectly rational expected utility maximisers. It gives good predictions - but it remains a model, with a domain of validity. You wouldn't conclude from that economic model that, say, mental illnesses don't exist. Similarly, modelling each life as having the same value and maximising expected lives saved is sensible and intuitive in many scenarios - but not necessarily all.

Maybe if we had a bit more information about the affected populations, we could use a more sophisticated model, such as one incorporating quality adjusted life years (QALY). Or maybe we could let other factors affect our thinking - what if we had to choose between saving a population of 1000 versus a population of 1001, of same average QALYs, but where the first set contained the entire Awá tribe/culture of 300 people, and the second is made up of representatives from much larger ethnic groups, much more culturally replaceable? Should we let that influence our decision? Well maybe we should, maybe we shouldn't, but it would be wrong to say "well, I would really like to save the Awá, but the model I settled on earlier won't allow me to, so I best follow the model". The models are there precisely to model our moral intuitions (the clue is in the name), not freeze them.

## The repugnant conclusion is at the end of a flimsy chain

There is a seemingly sound argument for the repugnant conclusion, which goes some way towards making total utilitarianism plausible. It goes like this:

1. Start with a population of very happy/utilitied/welfared/preference satisfied people.
2. Add other people whose lives are worth living, but whose average "utility" is less than that of the initial population.
3. Redistribute "utility" in an egalitarian way across the whole population, increasing the average a little as you do so (but making sure the top rank have their utility lowered).
4. Repeat as often as required.
5. End up with a huge population whose lives are barely worth living.

If all these steps increase the quality of the outcome (and it seems intuitively that they do), then the end state much be better than the starting state, agreeing with total utilitarianism. So, what could go wrong with this reasoning? Well, as seen before, the term "utility" is very much undefined, as is its scale - hence egalitarian is extremely undefined. So this argument is not mathematically precise, its rigour is illusionary. And when you recast the argument in qualitative terms, as you must, it become much weaker.

Going through the iteration, there will come a point when the human world is going to lose its last anime, its last opera, its last copy of the Lord of the Rings, its last mathematics, its last online discussion board, its last football game - anything that might cause more-than-appropriate enjoyment. At that stage, would you be entirely sure that the loss was worthwhile, in exchange of a weakly defined "more equal" society? More to the point, would you be sure that when iterating this process billions of times, *every* redistribution will be an improvement? This is a conjunctive statement, so you have to be nearly entirely certain of every link in the chain, if you want to believe the outcome. And, to reiterate, these links cannot be reduced to simple mathematical statements - you have to be certain that each step is qualitatively better than the previous one.

And you also have to be certain that your theory does not allow path dependency. One can take the perfectly valid position that "If there were an existing poorer population, then the right thing to do would be to redistribute wealth, and thus lose the last copy of Akira. However, currently there is no existing poor population, hence I would oppose it coming into being, precisely because it would result in the lose of Akira." You can reject this type of reasoning, and a variety of others that block the repugnant conclusion at some stage of the chain (the Stanford Encyclopaedia of Philosophy has a good entry on the Repugnant Conclusion and the arguments surrounding it). But most reasons for doing so already pre-suppose total utilitarianism. In that case, you cannot use the above as an argument for your theory.

## Hypothetical beings have hypothetical (and complicated) things say to you

There is another major strand of argument for total utilitarianism, which claims that we owe it to non-existent beings to satisfy their preferences, that they would prefer to exist rather than remain non-existent, and hence we should bring them into existence. How does this argument fare?

First of all, it should be emphasised that one is free to accept or reject that argument without any fear of inconsistency. If one maintains that never-existent beings have no relevant preferences, then one will never stumble over a problem. They don't exist, they can't make decisions, they can't contradict anything. In order to raise them to the point where their decisions are relevant, one has to raise them to existence, in reality or in simulation. By the time they can answer "would you like to exist?", they already do, so you are talking about whether or not to kill them, not whether or not to let them exist.

But secondly, it seems that the "non-existent beings" argument is often advanced for the sole purpose of arguing for total utilitarianism, rather than as a defensible position in its own right. Rarely are its implication analysed. What would a proper theory of non-existent beings look like?

Well, for a start the whole happiness/utility/preference problem comes back with extra sting. It's hard enough to make a utility function out of real world people, but how to do so with hypothetical people? Is it an essentially arbitrary process (dependent entirely on "which types of people we think of first"), or is it done properly, teasing out the "choices" and "life experiences" of the hypotheticals? In that last case, if we do it in too much detail, we could argue that we've already created the being in simulation, so it comes back to the death issue.

But imagine that we've somehow extracted a utility function from the preferences of non-existent beings. Apparently, they would prefer to exist rather than not exist. But is this true? There are many people in the world who would prefer not to commit suicide, but would not mind much if external events ended their lives - they cling to life as a habit. Presumably non-existent versions of them "would not mind" remaining non-existent.

Even for those that would prefer to exist, we can ask questions about the intensity of that desire, and how it compares with their other desires. For instance, among these hypothetical beings, some would be mothers of hypothetical infants, leaders of hypothetical religions, inmates of hypothetical prisons, and would only prefer to exist if they could bring/couldn't bring the rest of their hypothetical world with them. But this is ridiculous - we can't bring the hypothetical world with them, they would grow up in ours - so are we only really talking about the preferences of hypothetical babies, or hypothetical (and non-conscious) foetuses?

If we do look at adults, bracketing the issue above, then we get some that would prefer that they not exist, as long as certain others do - or conversely that they not exist, as long as others also not exist. How should we take that into account? Assuming the universe infinite, any hypothetical being would exist somewhere. Is mere existence enough, or do we have to have a large measure or density of existence? Do we need them to exist close to us? Are their own preferences relevant - ie we only have a duty to bring into the world, those beings that would desire to exist in multiple copies everywhere? Or do we feel these have already "enough existence" and select the under-counted beings? What if very few hypothetical beings are total utilitarians - is that relevant?

On a more personal note, every time we make a decision, we eliminate a particular being. We can not longer be the person who took the other job offer, or read the other book at that time and place. As these differences accumulate, we diverge quite a bit from what we could have been. When we do so, do we feel that we're killing off these extra hypothetical beings? Why not? Should we be compelled to lead double lives, assuming two (or more) completely separate identities, to increase the number of beings in the world? If not, why not?

These are some of the questions that a theory of non-existent beings would have to grapple with, before it can become an "obvious" argument for total utilitarianism.

# Moral uncertainty: total utilitarianism doesn't win by default

An argument that I have met occasionally is that while other ethical theories such as average utilitarianism, birth-death asymmetry, path dependence, preferences of non-loss of culture, etc... may have some validity, total utilitarianism wins as the population increases because the others don't scale in the same way. By the time we reach the trillion trillion trillion mark, total utilitarianism will completely dominate, even if we gave it little weight at the beginning.

But this is the wrong way to compare competing moral theories. Just as different people's utilities don't have a common scale, different moral utilities don't have a common scale. For instance, would you say that square-total utilitarianism is certainly wrong? This theory is simply total utilitarianism further multiplied by the population; it would correspond roughly to the number of connections between people. Or what about exponential-square-total utilitarianism? This would correspond roughly to the set of possible connections between people. As long as we think that exponential-square-total utilitarianism is not certainly completely wrong, then the same argument as above would show it quickly dominating as population increases.

Or what about 3^^^3 average utilitarianism - which is simply average utilitarianism, multiplied by 3^^^3? Obviously that example is silly - we know that rescaling shouldn't change anything about the theory. But similarly, dividing total utilitarianism by 3^^^3 shouldn't change anything, so total utilitarianism's scaling advantage is illusory.

As mentioned before, comparing different utility functions is a hard and subtle process. One method that seems to have surprisingly nice properties (to such an extent that I recommend always using as a first try) is to normalise the lowest possible attainable utility to zero, the highest attainable utility to one, multiply by the weight you give to the theory, and then add the normalised utilities together.

For instance, assume you equally valued average utilitarianism and total utilitarianism, giving them both weights of one (and you had solved all the definitional problems above). Among the choices you were facing, the worst outcome for both theories is an empty world. The best outcome for average utilitarianism would be ten people with an average "utility" of 100. The best outcome for total utilitarianism would be a quadrillion people with an average "utility" of 1. Then how would either of those compare to ten trillion people with an average utility of 60? Well, the normalised utility of this for the average utilitarian is 0.6, while for the total utilitarian it's also 60/100=0.6, and 0.6+0.6=1.2. This is better that the utility for the small world ($1+10^{-9}$) or the large world (0.01+1), so it beats either of the extremal choices.

Extending this method, we can bring in such theories as exponential-square-total utilitarianism (probably with small weights!), without needing to fear that it will swamp all other moral theories. And with this normalisation (or similar ones), even small weights to moral theories such as "culture has some intrinsic value" will often prevent total utilitarianism from walking away with *all* of the marbles.

# (Population) ethics is still hard

What is the conclusion? At Less Wrong, we're used to realising that ethics is hard, that value is fragile, that there is no single easy moral theory to safely program the AI with. But it seemed for a while that population ethics might be different - that there may be natural and easy ways to determine what to do with large groups, even though we couldn't decide what to do with individuals. I've argued strongly here that it's not the case - that population ethics remain hard, that we have to figure out what theory we want to have without access to easy shortcuts.

But in another way it's liberating. To those who are mainly total utilitarians but internally doubt that a world with infinitely many barely happy people surrounded by nothing but "muzak and potatoes" is really among the best of the best - well, you don't have to convince yourself of that. You may choose to believe it, or you may choose not to. No voice in the sky or in the math will force you either way. You can start putting together a moral theory that incorporates *all* your moral intuitions - those that drove you to total utilitarianism, and those that don't quite fit in that framework.

# Glenn Beck discusses the Singularity, cites SI researchers

From the final chapter of his new book *Cowards*, titled "Adapt or Die: The Coming Intelligence Explosion."

> The year is 1678 and you've just arrived in England via a time machine. You take out your new iPhone in front of a group of scientists who have gathered to marvel at your arrival.
>
> "Siri," you say, addressing the phone's voice-activated artificial intelligence system, "play me some Beethoven."
>
> *Dunh-Dunh-Dunh-Duuunnnhhh!* The famous opening notes of Beethoven's Fifth Symphony, stored in your music library, play loudly.
>
> "Siri, call my mother."
>
> Your mother's face appears on the screen, a Hawaiian beach behind her. "Hi, Mom!" you say. "How many fingers am I holding up?"
>
> "Three," she correctly answers. "Why haven't you called more—"
>
> "Thanks, Mom! Gotta run!" you interrupt, hanging up.
>
> "Now," you say. "Watch this."
>
> Your new friends look at the iPhone expectantly.
>
> "Siri, I need to hide a body."
>
> Without hesitation, Siri asks: "What kind of place are you looking for? Mines, reservoirs, metal foundries, dumps, or swamps?" (I'm not kidding. If you have an iPhone 4S, try it.)
>
> You respond "Swamps," and Siri pulls up a satellite map showing you nearby swamps.
>
> The scientists are shocked into silence. *What is this thing that plays music, instantly teleports video of someone across the globe, helps you get away with murder, and is small enough to fit into a pocket?*
>
> At best, your seventeenth-century friends would worship you as a messenger of God. At worst, you'd be burned at the stake for witchcraft. After all, as science fiction author Arthur C. Clarke once said, "Any sufficiently advanced technology is indistinguishable from magic."
>
> Now, imagine telling this group that capitalism and representative democracy will take the world by storm, lifting hundreds of millions of people out of poverty. Imagine telling them their descendants will eradicate smallpox and regularly live seventy-five or more years. Imagine telling them that men will walk on the moon, that planes, flying hundreds of miles an hour, will transport people around the

world, or that cities will be filled with buildings reaching thousands of feet into the air.

They'd probably escort you to the madhouse.

Unless, that is, one of the people in that group had been a man named Ray Kurzweil.

Kurzweil is an inventor and futurist who has done a better job than most at predicting the future. Dozens of the predictions from his 1990 book *The Age of Intelligent Machines* came true during the 1990s and 2000s. His follow-up book, *The Age of Spiritual Machines*, published in 1999, fared even better. Of the 147 predictions that Kurzweil made for 2009, 78 percent turned out to be entirely correct, and another 8 percent were roughly correct. For example, even though every portable computer had a keyboard in 1999, Kurzweil predicted that most portable computers would lack a keyboard by 2009. It turns out he was right: by 2009, most portable computers were MP3 players, smartphones, tablets, portable game machines, and other devices that lacked keyboards.

Kurzweil is most famous for his "law of accelerating returns," the idea that technological progress is generally "exponential" (like a hockey stick, curving up sharply) rather than "linear" (like a straight line, rising slowly). In nongeek-speak that means that our knowledge is like the compound interest you get on your bank account: it increases exponentially as time goes on because it keeps building on itself. We won't experience one hundred years of progress in the twenty-first century, but rather twenty thousand years of progress (measured at today's rate).

Many experts have criticized Kurzweil's forecasting methods, but a careful and extensive review of technological trends by researchers at the Santa Fe Institute came to the same basic conclusion: technological progress generally tends to be exponential (or even faster than exponential), not linear.

So, what does this mean? In his 2005 book *The Singularity Is Near*, Kurzweil shares his predictions for the next few decades:

- In our current decade, Kurzweil expects real-time translation tools and automatic house-cleaning robots to become common.
- In the 2020s he expects to see the invention of tiny robots that can be injected into our bodies to intelligently find and repair damage and cure infections.
- By the 2030s he expects "mind uploading" to be possible, meaning that your memories and personality and consciousness could be copied to a machine. You could then make backup copies of yourself, and achieve a kind of technological immortality.

[sidebar]

*Age of the Machines?*

"We became the dominant species on this planet by being the most intelligent species around. This century we are going to cede that crown to machines. After we do that, it will be them steering history rather than us."

—Jaan Tallinn, co-creator of Skype and Kazaa

[/sidebar]

If any of that sounds absurd, remember again how absurd the eradication of smallpox or the iPhone 4S would have seemed to those seventeenth-century scientists. That's because the human brain is conditioned to believe that the past is a great predictor of the future. While that might work fine in some areas, technology is not one of them. Just because it took decades to put two hundred transistors onto a computer chip doesn't mean that it will take decades to get to four hundred. In fact, Moore's Law, which states (roughly) that computing power doubles every two years, shows how technological progress must be thought of in terms of "hockey stick" progress, not "straight line" progress. Moore's Law has held for more than half a century already (we can currently fit 2.6 billion transistors onto a single chip) and there's little reason to expect that it won't continue to.

But the aspect of his book that has the most far-ranging ramifications for us is Kurzweil's prediction that we will achieve a "technological singularity" in 2045. He defines this term rather vaguely as "a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed."

Part of what Kurzweil is talking about is based on an older, more precise notion of "technological singularity" called an *intelligence explosion*. An intelligence explosion is what happens when we create artificial intelligence (AI) that is better than we are *at the task of designing artificial intelligences*. If the AI we create can improve its own intelligence without waiting for humans to make the next innovation, this will make it even more capable of improving its intelligence, which will . . . well, you get the point. The AI can, with enough improvements, make itself smarter than all of us mere humans put together.

The really exciting part (or the scary part, if your vision of the future is more like the movie The Terminator) is that, once the intelligence explosion happens, we'll get an AI that is as superior to us at science, politics, invention, and social skills as your computer's calculator is to you at arithmetic. The problems that have occupied mankind for decades— curing diseases, finding better energy sources, etc.— could, in many cases, be solved in a matter of weeks or months.

Again, this might sound far-fetched, but Ray Kurzweil isn't the only one who thinks an intelligence explosion could occur sometime this century. Justin Rattner, the chief technology officer at Intel, predicts some kind of Singularity by 2048. Michael Nielsen, co-author of the leading textbook on quantum computation, thinks there's a decent chance of an intelligence explosion by 2100. Richard Sutton, one of the biggest names in AI, predicts an intelligence explosion near the middle of the century. Leading philosopher David Chalmers is 50 percent confident an intelligence explosion will occur by 2100. Participants at a 2009 conference on AI tended to be 50 percent confident that an intelligence explosion would occur by 2045.

If we can properly prepare for the intelligence explosion and ensure that it goes well for humanity, it could be the best thing that has ever happened on this fragile planet. Consider the difference between humans and chimpanzees, which share

95 percent of their genetic code. A relatively small difference in intelligence gave humans the ability to invent farming, writing, science, democracy, capitalism, birth control, vaccines, space travel, and iPhones— all while chimpanzees kept flinging poo at each other.

[sidebar]

*Intelligent Design?*

The thought that machines could one day have superhuman abilities should make us nervous. Once the machines are smarter and more capable than we are, we won’t be able to negotiate with them any more than chimpanzees can negotiate with us. What if the machines don’t want the same things we do?

The truth, unfortunately, is that every kind of AI we know how to build today definitely would not want the same things we do. To build an AI that does, we would need a more flexible “decision theory” for AI design and new techniques for making sense of human preferences. I know that sounds kind of nerdy, but AIs are made of math and so math is really important for choosing which results you get from building an AI.

These are the kinds of research problems being tackled by the Singularity Institute in America and the Future of Humanity Institute in Great Britain. Unfortunately, our silly species still spends more money each year on lipstick research than we do on figuring out how to make sure that the most important event of this century (maybe of all human history)— the intelligence explosion— actually goes well for us.

[/sidebar]

Likewise, self-improving machines could perform scientific experiments and build new technologies much faster and more intelligently than humans can. Curing cancer, finding clean energy, and extending life expectancies would be child’s play for them. Imagine living out your own personal fantasy in a different virtual world every day. Imagine exploring the galaxy at near light speed, with a few backup copies of your mind safe at home on earth in case you run into an exploding supernova. Imagine a world where resources are harvested so efficiently that everyone’s basic needs are taken care of, and political and economic incentives are so intelligently fine-tuned that “world peace” becomes, for the first time ever, more than a Super Bowl halftime show slogan.

With self-improving AI we may be able to eradicate suffering and death just as we once eradicated smallpox. It is not the limits of nature that prevent us from doing this, but only the limits of our current understanding. It may sound like a paradox, but it’s our brains that prevent us from fully understanding our brains.

## Turf Wars

At this point you might be asking yourself: “Why is this topic in this book? What does any of this have to do with the economy or national security or politics?”

In fact, it has everything to do with all of those issues, plus a whole lot more. The intelligence explosion will bring about change on a scale and scope not seen in the history of the world. If we don’t prepare for it, things could get very bad, very fast. But if we do prepare for it, the intelligence explosion could be the best thing that has happened since . . . literally ever.

But before we get to the kind of life-altering progress that would come after the Singularity, we will first have to deal with a lot of smaller changes, many of which will throw entire industries and ways of life into turmoil. Take the music business, for example. It was not long ago that stores like Tower Records and Sam Goody were doing billions of dollars a year in compact disc sales; now people buy music from home via the Internet. Publishing is currently facing a similar upheaval. Newspapers and magazines have struggled to keep subscribers, booksellers like Borders have been forced into bankruptcy, and customers are forcing publishers to switch to ebooks faster than the publishers might like.

All of this is to say that some people are already witnessing the early stages of upheaval firsthand. But for everyone else, there is still a feeling that something is different this time; that all of those years of education and experience might be turned upside down in an instant. They might not be able to identify it exactly but they realize that the world they’ve known for forty, fifty, or sixty years is no longer the same.

There’s a good reason for that. We feel it and sense it because it’s true. It’s happening. There’s absolutely no question that the world in 2030 will be a very different place than the one we live in today. But there is a question, a large one, about whether that place will be better or worse.

It’s human nature to resist change. We worry about our families, our careers, and our bank accounts. The executives in industries that are already experiencing cataclysmic shifts would much prefer to go back to the way things were ten years ago, when people still bought music, magazines, and books in stores. The future was predictable. Humans like that; it’s part of our nature.

But predictability is no longer an option. The intelligence explosion, when it comes in earnest, is going to change everything— we can either be prepared for it and take advantage of it, or we can resist it and get run over.

Unfortunately, there are a good number of people who are going to resist it. Not only those in affected industries, but those who hold power at all levels. They see how technology is cutting out the middlemen, how people are becoming empowered, how bloggers can break national news and YouTube videos can create superstars.

And they don’t like it.

### A Battle for the Future

Power bases in business and politics that have been forged over decades, if not centuries, are being threatened with extinction, and they know it. So the owners of that power are trying to hold on. They think they can do that by dragging us backward. They think that, by growing the public's dependency on government, by taking away the entrepreneurial spirit and rewards and by limiting personal freedoms, they can slow down progress.

But they're wrong. The intelligence explosion is coming so long as science itself continues. Trying to put the genie back in the bottle by dragging us toward serfdom won't stop it and will, in fact, only leave the world with an economy and society that are completely unprepared for the amazing things that it could bring.

Robin Hanson, author of "The Economics of the Singularity" and an associate professor of economics at George Mason University, wrote that after the Singularity, "The world economy, which now doubles in 15 years or so, would soon double in somewhere from a week to a month."

That is unfathomable. But even if the rate were much slower, say a doubling of the world economy in two years, the shock-waves from that kind of growth would still change everything we've come to know and rely on. A machine could offer the ideal farming methods to double or triple crop production, but it can't force a farmer or an industry to implement them. A machine could find the cure for cancer, but it would be meaningless if the pharmaceutical industry or Food and Drug Administration refused to allow it. The machines won't be the problem; humans will be.

And that's why I wanted to write about this topic. We are at the forefront of something great, something that will make the Industrial Revolution look in comparison like a child discovering his hands. But we have to be prepared. We must be open to the changes that will come, because they will come. Only when we accept that will we be in a position to thrive. We can't allow politicians to blame progress for our problems. We can't allow entrenched bureaucrats and power-hungry executives to influence a future that they may have no place in.

Many people are afraid of these changes— of course they are: it's part of being human to fear the unknown— but we can't be so entrenched in the way the world works now that we are unable to handle change out of fear for what those changes might bring.

Change is going to be as much a part of our future as it has been of our past. Yes, it will happen faster and the changes themselves will be far more dramatic, but if we prepare for it, the change will mostly be positive. But that preparation is the key: we need to become more well-rounded as individuals so that we're able to constantly adapt to new ways of doing things. In the future, the way you do your job may change four to five or fifty times over the course of your life. Those who cannot, or will not, adapt will be left behind.

At the same time, the Singularity will give many more people the opportunity to be successful. Because things will change so rapidly there is a much greater likelihood that people will find something they excel at. But it could also mean that people's successes are much shorter-lived. The days of someone becoming a legend in any one business (think Clive Davis in music, Steven Spielberg in movies, or the Hearst family in publishing) are likely over. But those who embrace and adapt to the coming changes, and surround themselves with others who have done the same, will flourish.

> When major companies, set in their ways, try to convince us that change is bad and that we must stick to the status quo, no matter how much human inquisitiveness and ingenuity try to propel us forward, we must look past them. We must know in our hearts that these changes will come, and that if we welcome them into our world, we’ll become more successful, more free, and more full of light than we could have ever possibly imagined.
>
> Ray Kurzweil once wrote, “The Singularity is near.” The only question will be whether we are ready for it.

The citations for the chapter include:

- Luke Muehlhauser and Anna Salamon, "Intelligence Explosion: Evidence and Import"
- Daniel Dewey, "Learning What to Value"
- Eliezer Yudkowsky, "Artificial Intelligence as a Positive and a Negative Factor in Global Risk"
- Luke Muehlhauser and Louie Helm, "The Singularity and Machine Ethics"
- Luke Muehlhauser, "So You Want to Save the World"
- Michael Anissimov, "The Benefits of a Successful Singularity"

# Intellectual insularity and productivity

*Guys I'd like your opinion on something.*

**Do you think LessWrong is too intellectually insular?** What I mean by this is that we very seldom seem to adopt useful vocabulary or arguments or information from outside of LessWrong. For example all I can think of is some of Robin Hanson's and Paul Graham's stuff. But I don't think Robin Hanson really counts as Overcoming Bias used to be LessWrong.

The community seems to not update on ideas and concepts that didn't originate here. The only major examples fellow LWers brought up in conversation where works that Eliezer cited as great or influential. :/

***Edit:*** *Apparently this has been a source of much confusion and mistargeted replies. While I wouldn't mind even more references to quality outside writing, this wasn't my concern. I'm surprised this was problematic to understand for two reasons. First I gave examples of two thinker that aren't often linked to by recent articles on LW yet have clearly greatly influenced us. Secondly this is a trivially false interpretation, as my own submission history shows (it is littered with well received outside links). I think this arises because when I wrote "we seem to not* ***update*** *on ideas and concepts that didn't originate here" people read it as "we don't* ***link*** *to ideas and concepts" or maybe "we don't* ***talk*** *about ideas and concepts" from outside. I clarified this several times in the comments, most extensively here. Yet it doesn't seem to have made much of an impact. Maybe it will be easier to understand if I put it this way, interesting material from the outside never seems to get added to something like the sequences or the wiki. The sole exception to this is hunting even more academic references for the conclusions and concepts we already know and embrace. Thus while individuals will update on them and perhaps even reference them in the future* ***the community as a whole will not****. They don't become part of the expected background knowledge when discussing certain topics. Over time their impact thus fades in a way the old core material doesn't.*

Another thing, I could be wrong about this naturally, but it seems to clear that LessWrong has **not** grown. I'm not talking numerically. I can't put my finger to major progress done in the past 2 years. I have heard several other users express similar sentiments. To quote one user:

> I notice that, in topics that Eliezer did not *explicitly* cover in the sequences (and some that he did), LW has made zero progress in general.

I've recently come to think this is probably true to the first approximation. I was checking out a blogroll and saw LessWrong listed as Eliezer's blog about rationality. I realized that essentially *it is*. And worse this makes it a very crappy blog since the author doesn't make new updates any more. Originally the man had high hopes for the site. He wanted to build something that could keep going on its own, growing without him. It turned out to be a community mostly dedicated to studying the scrolls he left behind. We don't even seem to do a good job of getting others to *read* the scrolls.

Overall there seems to be little enthusiasm for actually systematically reading the old material. I'm going to share my take on what is I think a symptom of this. I was debating which title to pick for my first ever original content Main article (it was originally titled "On Conspiracy Theories") and made what at first felt like a joke but then took on a horrible ring of:

> **Over time the meaning of an article will tend to converge with the literal meaning of its title.**

We like linking articles, and while people may read a link the first time, they don't tend to read it the second or third time they run across it. The phrase is eventually picked up and used out the appropriate of context. Something that was supposed to be shorthand for a nuanced argument starts to mean *exactly* what "it says". Well not *exactly*, people still recall it is a vague applause light. Which is actually *worse*.

I cited precisely "Politics is the Mindkiller" as an example of this. In the original article Eliezer basically argues that gratuitous politics, political thinking that isn't outweighed by its value to the art of rationality, is to be avoided. This soon came to meant it is forbidden to discuss politics in Main and Discussion articles, though it does live in the comment sections.

**Now the question if LessWrong remains productive intellectually, is separate from the question of it being insular. But I feel both need to be discussed.** If our community wasn't growing and it wasn't insular either, it could at least remain relevant.

This site has a wonderful ethos for discussion and thought. Why do we seem to be wasting it?

# How to Run a Successful Less Wrong Meetup

Always wanted to run a Less Wrong meetup, but been unsure of how? The *How to Run a Successful Less Wrong Meetup* booklet is here to help you!

The 33-page document draws from consultations with more than a dozen Less Wrong meetup group organizers. Stanislaw Boboryk created the document design. Luke provided direction, feedback, and initial research, and I did almost all the writing.

The booklet starts by providing some motivational suggestions on *why* you'd want to create a meetup in the first place, and then moves on to the subject of organizing your first one. Basics such as choosing a venue, making an announcement, and finding something to talk about once at the meetup, are all covered. This section also discusses pioneering meetups in foreign cities and restarting inactive meetup groups.

For those who have already established a meetup group, the booklet offers suggestions on things such as attracting new members, maintaining a pleasant atmosphere, and dealing with conflicts within the group. The "How to Build Your Team of Heroes" section explains the roles that are useful for a meetup group to fill, ranging from visionaries to organizers.

If you're unsure of what exactly to *do* at meetups, the guide describes many options, from different types of discussions to nearly 20 different games and exercises. All the talk and philosophizing in the world won't do much good if you don't actually *do* things, so the booklet also discusses long-term projects that you can undertake. Some people attend meetups to just have fun and to be social, and others to improve themselves and the world. The booklet has been written to be useful for both kinds of people.

In order to inspire you and let you see what others have done, the booklet also has brief case studies and examples from real meetup groups around the world. You can find these sprinkled throughout the guide.

**This is just the first version of the guide.** We *will* continue working on it. If you find mistakes, or think that something is unclear, or would like to see some part expanded, or if you've got good advice you think should be included... please let me know! You can contact me at kaj.sotala@intelligence.org.

A large number of people have helped in various ways, and I hope that I've remembered to mention most of them in the acknowledgements. If you've contributed to the document but don't see your name mentioned, please send me a message and I'll have that fixed!

The booklet has been illustrated with pictures from various meetup groups. Meetup organizers sent me the pictures for this use, and I explicitly asked them to make sure that everyone in the photos was fine with it. Regardless, if there's a picture that you find objectionable, please contact me and I'll have it replaced with something else.

# [Link] Can We Reverse The Stanford Prison Experiment?

From the Harvard Business Review, an article entitled: "Can We Reverse The Stanford Prison Experiment?"

By: Greg McKeown
Posted: June 12, 2012

Clicky Link of Awesome! Wheee! Push me!

**Summary**:

### Royal Canadian Mounted Police attempt a program where they hand out "Positive Tickets"

> Their approach was to try to catch youth doing the right things and give them a Positive Ticket. The ticket granted the recipient free entry to the movies or to a local youth center. They gave out an average of 40,000 tickets per year. That is three times the number of negative tickets over the same period. As it turns out, and unbeknownst to Clapham, that ratio (2.9 positive affects to 1 negative affect, to be precise) is called the Losada Line. It is the minimum ratio of positive to negatives that has to exist for a team to flourish. On higher-performing teams (and marriages for that matter) the ratio jumps to 5:1. But does it hold true in policing?
>
> According to Clapham, youth recidivism was reduced from 60% to 8%. Overall crime was reduced by 40%. Youth crime was cut in half. And it cost one-tenth of the traditional judicial system.

### This idea can be applied to Real Life

> The lesson here is to create a culture that *immediately* and sincerely celebrates victories. Here are three simple ways to begin:
>
> **1. Start your next staff meeting with five minutes on the question: "What has gone right since our last meeting?"** Have each person acknowledge someone else's achievement in a concrete, sincere way. Done right, this very small question can begin to shift the conversation.
>
> **2. Take two minutes every day to try to catch someone doing the right thing.** It is the fastest and most positive way for the people around you to learn when they are getting it right.
>
> **3. Create a virtual community board where employees, partners and even customers can share what they are grateful for daily.** Sounds idealistic? Vishen Lakhiani, CEO of Mind Valley, a new generation media and publishing company, has done just that at Gratitude Log. (Watch him explain how it works here).

# Introduction to Game Theory: Sequence Guide

This sequence of posts is a primer on game theory intended at an introductory level. Because it is introductory, Less Wrong veterans may find some parts boring, obvious, or simplistic - although hopefully nothing is so simplistic as to be outright wrong.

Parts of this sequence draw heavily upon material from *The Art of Strategy* by Avinash Dixit and Barry Nalebuff, and it may in part be considered a (very favorable) review of the book accompanied by an exploration of its content. I have tried to include enough material to be useful, but not so much material that it becomes a plagiarism rather than a review (it's probably a bad idea to pick a legal fight with people who write books called *The Art of Strategy*.) Therefore, for the most complete and engaging presentation of this material, I highly recommend the original book.

All posts will be linked from here as they go up:

1. Introduction to Game Theory: Sequence Guide
2. Backward Reasoning Over Decision Trees
3. Nash Equilibria and Schelling Points
4. Introduction to Prisoners' Dilemma
5. Real World Solutions to Prisoners' Dilemmas
6. Interlude for Behavioral Economics
7. What Is Signaling, Really?
8. Bargaining and Auctions
9. Imperfect Voting Systems
10. Game Theory As A Dark Art

Special thanks to Luke for his book recommendation and his strong encouragement to write this.

# Conspiracy Theories as Agency Fictions

**Related to:** Consider Conspiracies, What causes people to believe in conspiracy theories?

*Here I consider in some detail a failure mode that classical rationality often recognizes. Unfortunately nearly all heuristics normally used to detect it seem remarkably vulnerable to misfiring or being exploited by others. I advocate an approach where we try our best to account for the key bias, seeing agency where there is none, while trying to minimize the risk of being tricked into dismissing claims because of boo lights.*

## What does calling something a "conspiracy theory" tell us?

What is a conspiracy theory? Explanations that invoke plots orchestrated by covert groups are easily called or thought of as such. In a more legal sense conspiracy is an agreement between persons to mislead or defraud others. This simple story gets complicated because people aren't very clear on what they consider a conspiracy.

To give an example, is explicit negotiation or agreement really necessary to call something a conspiracy? Does *silent* cooperation on Prisoner's Dilemma count? What if the players are deceiving themselves that they are really following a different goal and the resulting cooperation is just a side effect? How could we tell the difference and would it matter? The latter is especially interesting if one applies the anthropic principle to social attitudes and norms.

The phrase is also **a convenient tool** to mark an opponent's tale as low status and unworthy of further investigation. A boo light easily applied to anything that has people acting in something that can be framed as self-interest and happens to be few inferential jumps away from the audience. Not only is its use in this way well known, this is arguably the *primary meaning* of calling an argument a conspiracy theory.

We have *plenty* of historical examples of high-stakes conspiracies so we know **they *can* be the right answer**. Noting this and putting aside the misuse of the label, people *do* engage in crafting conspiracy theories when they just aren't needed. Entire communities can fixate on them or fail to call such bad thinking out. Why does this happen? Humans being the social animals that we are, the group dynamics at work probably need an article or sequence of their own. It should suffice for now to point to belief as attire, the bandwagon effect and Robin Hanson's take on status. **Let's rather consider the question of why individuals may be biased towards such explanations.** Why do they privilege the hypothesis?

## When do they seem more likely than they are?

First off we have a hard time understanding that coordination is hard. Seeing a large pay off available and thinking it easily in reach if "we could just get along" seems like a classical failing. Our pro-social sentiments lead us to downplay such barriers in our future plans. Motivated cognition on behalf of assessing the threat potential of perceived *enemies* or *strangers* likely shares this problem. Even if we avoid this, we may still be lost since the second big relevant thing is our tendency for anthropomorphizing things that better not be. Ours is a paranoid brain seeing agency in every shadow or strange sound. The cost of false positives was once reasonably low, while the cost of a false negative very high.

Our minds are also just plain lazy. We are pretty good at modelling other human minds and considering just how hard the task really is, we do a pretty remarkable job of it. If you are stuck in relative ignorance on a subject, say the weather, dancing to appease the sky spirits makes sense. After all the weather is pretty capricious and angry sky spirits is a model that makes as much or more sense as any other model you know. Unlike some other models this one is at least cheap to run on your brain! The modern world is remarkably complex. Do we see ghosts in it?

Our Dunbarian minds probably just plain can't *get* how a society can be that complex and unpredictable without it being "planned" by a cabal of Satan or Heterosexual White Males or the Illuminati (but I repeat myself twice) scheming to make weird things happen in our oblivious small stone age tribe. **Learning about useful models helps people escape anthropomorphizing human society or the economy or government**. The latter is particularly salient. I think most people slip up occasionally in assuming that say something like the United States government can be successfully modelled as a single agent to explain most of its "actions". To make matters worse it is a common literary device used by pundits.

A mysterious malignant agency or someone keeping a secret playing the role of the villain makes a good story. *Humans love stories*. Its *fun* to think in stories. Any real conspiracy revealed will probably be widely publicized. Peter Knight in his 2003 book cites historians who have put forward the idea, that the United States is something of a home for popular conspiracy theories *because so many high-level ones have been undertaken and uncovered since the 1960s*. We are more likely to *hear* about real confirmed conspiracies today than ever before.

Wishful thinking also plays a role. **A universe where bad things happen because bad people make them to is appealing.** Getting rid of bad people, even very bad people, is easy compared to all the different things one has to do to make sure bad things don't happen in a universe that doesn't care about us and where really bad things are allowed to happen. Finding bad people whether there are or aren't is a problematic tendency. The sad thing is that this may also be how we often manage to coordinate. Do all theories of legitimacy also perhaps rest on the same cognitive failings that conspiracy theories do? The difference between a shadowy cabal we need to get rid of and an institution worthy of respect may be just some bad luck.

## How this misleads us

Putting aside such wild speculation, what should we take away from this? When do conspiracy theories seem more likely than they are?

- The phenomena is unpredictable or can't be modelled very well
- Models used by others are hard to understand or are very counter-intuitive
- Thinking about the subject significantly strains cognitive resources
- The theory explains why bad things happen or why something went wrong
- The theory requires coordination

When you see these features you probably find the theory more plausible than it is.

But how many here are likely to accept "conspiracy theories"? To do so with stuff that actually gets called a conspiracy theory doesn't fit our tribal attire. Reverse stupidity may be particularly problematic for us on this topic. *Being open to thinking conspiracy is recommended.* Just remember to compare how probable it is in relation to other explanations. **It is important to call out people who misuse the tag for rhetorical gain.**

This applies to debunking as well. Don't go wildly contrarian. But remember that even things that are tagged conspiracy theories are surprisingly popular. How popular might false theories that avoid that tag be? History shows us we don't have the luxury of hoping that kind of thing just doesn't happen in human societies. **When assessing an explanation sharing the key features that make conspiracy theories seem more plausible than they are, compensate as you would with a conspiracy theory.**

But don't listen to me, I'm talking conspiracy theories.

---

**Note:** This article started out as a public draft, feedback to other such drafts is always welcomed.  Special thanks to user **Villiam_Bur** for his commentary and user **copt** for proofreading and suggestions. Also thanks to the LessWrong IRC chatroom for last minute corrections and stylistic tips.

# Blogs by LWers

**Related to:** Wikifying the Blog List

LessWrong posters and readers are generally pretty cool people. Maybe they are interesting bloggers too. And I'm not just talking about rationalist material, that we'd ideally like to be cross posted on LessWrong, no gardening blogs are also fair game. I'm making this a discussion level post so more people can see the list. Please share links to blogs by former or current LWers. Surely the authors wouldn't mind, who wouldn't like more readers? Original list here.

**Anyone who wants to suggest a new blog for the list please follow this link.**

**Blogs by LWers:**

- RobinHanson --- Overcoming Bias (Katja Grace and Robert Wiblin post here as well)
- Katja Grace --- Meteuphoric (very cool old posts and summaries)
- muflax --- muflax' mindstream, daily
- TGGP --- Entitled To An Opinion
- Yvain --- Jackdaws love my big sphinx of quartz
- juliawise --- Giving Gladly, Radiant Things
- James_G --- Writings
- steven0461 --- Black Belt Bayesian
- James Miller --- Singluarity Notes
- Jsalvati --- Good Morning, Economics
- Will Newsome --- Computational Theology
- clarissethorn --- Clarrise Thorn
- Zack M. Davis --- An Algorithmic Lucidity
- Kaj_Sotala --- A view to the gallery of my mind
- SilasBarta --- Setting Things Straight
- tommcabe --- The Rationalist Conspiracy
- Alicorn --- Irregular Updates By An Irregular Person
- MBlume --- Baby, check this out; I've got something to say.
- ciphergoth --- Paul Crowley's blog (mostly about cryonics), Paul Crowley
- XiXiDu --- Alexander Kruel
- Aurini --- Stares At The World
- jkaufman --- Jeff Kaufman
- Bill_McGrath --- billmcgrathmusic
- Sister Y --- the view from hell
- PaulWright --- Paul Wright's blog
- _ozymandias --- http://ozyfrantz.com/
- mstevens --- stdout
- HughRistik --- Feminist Critics
- Julia_Galef --- Measure of Doubt
- NancyLebovitz --- Input Junkie
- David Gerard --- a bunch of them
- Jayson_Virissimo --- Jay, Quantified
- kpreid --- Kevin Reid's blog
- hegemonicon --- Coarse Grained
- Villiam_Bur --- bur.sk
- Emile --- The Rational Parent

- lukeprog --- Common Sense Atheism
- Grognor --- Grognor's Blog
- CarlShulman --- Reflective Disequilibrium
- OrphanWilde --- Aretae
- Alexei --- Bent Spoon Games Blog
- TimS --- Georgia Special Education Law Blog
- loup-valliant --- @ Loup's
- RolfAndreaseen --- Yngling Saga
- arundelo --- Aaron Brown
- peter_hurtford --- Greatplay.net
- brilee --- Modern Descartes
- gwern --- gwern.net
- erratio --- The merry-go-round of life
- jimmy --- The Art and Science of Cognitive Engineering
- alexvermeer --- alexvermeer.com
- sark --- sarkology
- gjm --- Scribble, scribble, scribble
- Giles --- Prince Mm Mm
- Chris Hallquist --- The Uncredible Hallq
- EricHerboso --- EricHerboso.org
- Eneasz - Death Is Bad
- Tuxedage - Essays and other Musings
- Federico - studiolo
- Trevor Blake - OVO, editor-Dora Marsden, lead judge-George Walford International Essay Prize
- Pablo_Stafforini -- Pablo's miscellany

## List of LWers on Twitter

**Note:** Anyone just digging for interesting blogs they would like to read but dosen't care if they are written by LWers or not should check out this thread or maybe this one. Did you guys know we have a wiki article with external resources? We do. Check that out as well. Maybe once we figure out which LWer blogs related to rationality on this list are particularly good we can add a few of them there too.

# [Link] The Greek Heliocentric Theory

*Summary: The Greeks likely rejected a heliocentric theory because it would conflict with the lack of any visible stellar parallax, not for egotistical, common-sense, or aesthetic reasons.*

I had always heard that the Greeks embraced a geocentric universe for common-sense, aesthetic reasons - not scientific ones. But it seems as if the real story is more complicated than that:

From Isomorphismes:

> Now this is the kicker in your Popperian dirtsack. The Greeks had the right theory (heliocentric solar system) but discarded it on the basis of experimental evidence! Never preach to me about progress-in-science when all you've heard is a one-liner about Popper and the communal acceptance of general relativity. Especially don't follow it up by saying that science marches toward the Truth whilst religion thwarts its progress. According to Astronomer Lisa, it's not true that the Greeks simply thought they and their Gods were at the centre of the Universe because they were egotistical. They reasoned to the geocentric conclusion based on quantitative evidence. How? They measured parallax.(Difference in stellar appearance from spring to fall, when we're on opposite sides of the Sun.) Given the insensitivity of their measurement tools at the time, the stars didn't change positions at all when the Earth moved to the other side of the Sun. Based on that, they rejected the heliocentric hypothesis. If the Earth actually did move around the Sun, then the stars would logically have to appear different from one time to another. But they remain ever fixed in the same place in the Heavens, therefore the Earth must be still (geocentric).

I dug a little bit deeper, and this seems to be more or less accurate. From The Greek Heliocentric Theory and its Abandonment:

> This paper then examines possible reasons for the Greek abandonment of the heliocentric theory and concludes that there is no reason to deplore its abandonment. In developing the heliocentric theory the Greeks had run the gamut of theorizing. We are indebted to the Alexandrians and Hipparchus for turning away from speculation to take up the recording of precise astronomical data. Here was laid the foundation upon which modern astronomy was built.
>
> Let us now suppose that Aristarchus' theory was widely circulated and that it was given careful consideration by leading astronomers. There is one objection that immediately arises when the earth is put in motion, the very difficulty which must have disquieted Copernicus and which caused Tycho Brahe shortly afterwards to renounce Copernicus' heliocentric system and to put the earth again at rest. (Tycho reverted to a system first suggested by some ancient Greek, who made the planets revolve about the sun and the sun about the earth.) The difficulty is this. As soon as the earth is set in motion in an annual revolution about the sun, the distance between any two of the earth's positions that are six months apart will be twice as great as the earth's distance from the sun. Over such vast distances some displacement in the positions of the stars ought to be observed. The more accurate the astronomical instruments and the greater the estimated distance of the sun, the more reason should there be to

> expect stellar displacement. Now it so happened that Aristarchus reached his conclusions at the very time when interest was keen at Alexandria and elsewhere in the Greek world in accurate observations and when marked improvements were being made in precision instruments. To appreciate these developments we need only recall the careful stellar catalogues of Aristyllus and Timocharis early in the third century B.C., the work of the latter enabling Hipparchus to discover the precession of the equinoxes, and the armillary sphere of Eratosthenes by which he was able to determine the obliquity of the ecliptic and the circumference of the earth. Hipparchus continued to make improvements in the next century. He, as we shall see, had a much better appreciation of the sun's great distance than Copernicus. Of course it was impossible to observe stellar displacement without the aid of a telescope. Inability to observe it left astronomers with only two alternatives: either the stars were so remote that it was impossible to detect displacement, or the earth would have to remain at rest.
>
> ..Heath was of the opinion that Hipparchus was responsible for the death of Aristarchus' theory, that the adherence of so preeminent an astronomer to a geocentric orientation sealed the doom of the heliocentric theory. This is a reasonable conjecture. Hipparchus was noted for his careful observations, his stellar catalogues, and the remarkable precision of his recordings of solar and lunar motions. According to Ptolemy he was devoted to truth above all else and because he did not possess sufficient data, he refused to attempt to account for planetary motions as he had for those of the sun and moon. His discovery of the precession of the equinoxes attests to the keenness of his observations. He came much closer to appreciating the vast distance of the sun than Copernicus did.
>
> ..We do not know whether or not Hipparchus ever seriously entertained Aristarchus' views about the earth's motions, but from what we have seen of his cautious and accurate methods, it is likely that he would have quickly rejected the heliocentric theory in the absence of visible stellar displacement.

And from The Ancient Greek Astronomers: A Remarkable Record of Ingenuity:

> Aristarchus was successful in explaining variations in brilliance and reverse courses of the planets, but planetary motions are far more complicated than that. Kepler was the first to realize that the planets do not describe circular orbits, but rather ellipses, and that the sun is not in the middle of these orbits but in the foci of the ellipses. That something was wrong might have been suspected as early as 330 B.C., for Callippus noticed that the seasons were not of the same length. He estimated their lengths between solstices and equinoxes to be 94, 92, 89, and 90 days- figures that are very nearly correct. Or to show the irregularities that might result from combining the eccentricities of the orbits of two planets, in some years Mars and the earth at closest approximation are 36 million miles apart and in other years (as in 1948) may be 63 million miles apart at their nearest approach. Now the Alexandrians were pointing their precision sights at the planets and must have been disturbed by these peculiarities. Furthermore they would have been less kindly disposed towards Aristarchus' explanation of the absence of visible stellar parallax by placing the stars at an almost infinite distance away because they had a better appreciation of the sun's vast distance and consequently would have stronger reason to expect to find parallax. It would seem that the more precise the instruments, the less likelihood there would be of the earth's being in motion.

# Bounded versions of Gödel's and Löb's theorems

While writing up decision theory results, I had to work out bounded versions of Gödel's second incompleteness theorem and Löb's theorem. Posting the tentative proofs here to let people find any major errors before adding formality. Any comments are welcome!

### Bounded Gödel's theorem

Informal meaning: if a theory T has a short proof that any proof of T's inconsistency must be long, then T is inconsistent.

Formal statement: there exists a total computable function f such that for any effectively generated theory T that includes Peano arithmetic, and for any integer n, if T proves in n symbols that "T can't prove a falsehood in f(n) symbols", then T is inconsistent.

### Proof

If the theory T takes more than n symbols to describe, the condition of the theorem can't hold, so let's assume T is smaller than that.

Let's take g(n)>>n and f(n)>>exp(g(n)). For any integer n, let's define a program R that enumerates all proofs in the theory T up to length g(n), looking for either a proof that R prints something (in which case R immediately halts without printing anything), or a proof that R doesn't print anything (in which case R prints something and halts). If no proof is found, R halts without printing anything.

Assume that the theory T is consistent. Note that if the program R finds some proof, then it makes the proved statement false. Also note that the theory T can completely simulate the execution of the program R in exp(g(n)) symbols. Therefore if the program R finds any proof, the resulting contradiction in the theory T can be "exposed" using exp(g(n)) symbols. Since f(n)>>exp(g(n)), and the theory T proves in n symbols that "T can't prove a falsehood in f(n) symbols", we see that T proves in slightly more than n symbols that the program R won't find any proofs. Therefore the theory T proves in slightly more than n symbols that the program R won't print anything. Since g(n)>>n, the program R will find that proof and print something, making the theory T inconsistent. QED.

(The main idea of the proof is taken from Ron Maimon, originally due to Rosser, and adapted to the bounded case.)

### Bounded Löb's theorem

Informal meaning: if having a bounded-size proof of statement S would imply the truth of S, and that fact has a short proof, then S is provable.

Formal statement: there exists a total computable function f such that for any effectively generated theory T that includes Peano arithmetic, any statement S in that theory, and any integer n, if T proves in n symbols that "if T proves S in f(n) symbols, then S is true", then T proves S.

**Proof**

Taking the contrapositive, the theory T proves in n symbols that "if the statement S is false, then T doesn't prove S in f(n) symbols". Therefore the theory T+¬S proves in n symbols that "T+¬S doesn't prove a falsehood in f(n) symbols" (I'm glossing over the tiny changes to n and f(n) here). Taking the function f from the bounded Gödel's theorem, we obtain that the theory T+¬S is inconsistent. That means T proves S. QED.

(The main idea of the proof is taken from Scott Aaronson, originally due to Kripke or Kreisel.)

As a sanity check, the regular versions of the theorems seem to be easy corollaries of the bounded versions. But of course there could still be errors...