

Best of LessWrong: September 2013

1. [Instinctive Frequentists, the Outside View, and de-Biasing](#)
2. [Polyphasic Sleep Seed Study: Reprise](#)
3. [Three ways CFAR has changed my view of rationality](#)
4. [Inferential silence](#)
5. [The genie knows, but doesn't care](#)
6. [Cooperating with agents with different ideas of fairness, while resisting exploitation](#)
7. [Notes on Brainwashing & 'Cults'](#)
8. [Making Fun of Things is Easy](#)
9. [A map of Bay Area memespace](#)
10. [I attempted the AI Box Experiment again! \(And won - Twice!\)](#)
11. [The Anti-Placebo Effect](#)
12. [I played the AI Box Experiment again! \(and lost both games\)](#)
13. [Probability, knowledge, and meta-probability](#)
14. [Book Review: Basic Category Theory for Computer Scientists \(MIRI course list\)](#)

Best of LessWrong: September 2013

1. [Instinctive Frequentists, the Outside View, and de-Biasing](#)
2. [Polyphasic Sleep Seed Study: Reprise](#)
3. [Three ways CFAR has changed my view of rationality](#)
4. [Inferential silence](#)
5. [The genie knows, but doesn't care](#)
6. [Cooperating with agents with different ideas of fairness, while resisting exploitation](#)
7. [Notes on Brainwashing & 'Cults'](#)
8. [Making Fun of Things is Easy](#)
9. [A map of Bay Area memespace](#)
10. [I attempted the AI Box Experiment again! \(And won - Twice!\)](#)
11. [The Anti-Placebo Effect](#)
12. [I played the AI Box Experiment again! \(and lost both games\)](#)
13. [Probability, knowledge, and meta-probability](#)
14. [Book Review: Basic Category Theory for Computer Scientists \(MIRI course list\)](#)

Instinctive Frequentists, the Outside View, and de-Biasing

In "[How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases](#)", [Gerd Gigerenzer](#) attempts to show that the whole "Heuristics and Biases" approach to analysing human reasoning is fundamentally flawed and incorrect.

In that he fails. His case depends on using the [frequentist](#) argument that probabilities cannot be assigned to single events or situations of subjective uncertainty, thus removing the possibility that people could be "wrong" in the scenarios where the biases were tested. (It is interesting to note that he ends up constructing "Probabilistic Mental Models", which are frequentist ways of assigning subjective probabilities - just as long as you don't call them that!).

But that dodge isn't sufficient. Take the famous example of the [conjunction fallacy](#), where most people are tricked to assigning a higher probability to "Linda is a bank teller AND is active in the feminist movement" than to "Linda is a bank teller". This error [persists](#) even when people take bets on the different outcomes. By betting more (or anything) on the first option, people are giving up free money. This is a failure of human reasoning, whatever one thinks about the morality of assigning probability to single events.

However, though the article fails to prove its case, it presents a lot of powerful results that may change how we think about biases. It presents weak evidence that people may be instinctive frequentist statisticians, and much stronger evidence that *many biases can go away when the problems are presented in frequentist ways*.

Now, it's known that people are more comfortable with frequencies than with probabilities. The examples in the paper extend that intuition. For instance, when people are asked:

There are 100 persons who fit the description above (i.e., Linda's). How many of them are:

- (a) bank tellers
- (b) bank tellers and active in the feminist movement.

Then the conjunction fallacy essentially disappears (22% of people make the error, rather than 85%). That is a huge difference.

Similarly, overconfidence. When people were 50 general knowledge questions and asked to rate their confidence for their answer on each question, they were systematically, massively overconfident. But when they were asked afterwards "How many of these 50 questions do you think you got right?", they were... underconfident. But only very slightly: they were essentially correct in their self-assessments. This can be seen as a use of the outside view - a use that is, in this case, entirely justified. People know their overall accuracy much better than they know their specific accuracy.

A more intriguing example makes the base-rate fallacy disappear. Presenting the problem in a frequentist way makes the fallacy vanish when computing false positives for tests on rare diseases - that's compatible with the general theme. But it really got interesting when people actively participated in the randomisation process. In the

standard problem, students were given thumbnail description of individuals, and asked to guess whether they were more likely to be engineers or lawyers. Half the time the students were told the descriptions were drawn at random from 30 lawyers and 70 engineers; the other half, the proportions were reversed. It turns out that students assigned similar guesses to lawyer and engineer in both setups, showing they were neglecting to use the 30/70 or 70/30 base-rate information.

Gigerenzer modified the setups by telling the students the 30/70 or 70/30 proportions and then having the students themselves draw each description (blindly) out of an urn before assessing it. In that case, base-rate neglect disappears.

Now, I don't find that revelation *quite* as superlatively exciting as Gigerenzer does. Having the students draw the description out of the urn is pretty close to whacking them on the head with the base-rate: it really focuses their attention on this aspect, and once it's risen to their attention, they're much more likely to make use of it. It's still very interesting, though, and suggests some practical ways of overcoming the base-rate problem that stop short of saying "hey, don't forget the base-rate".

There is a large literature out there [critiquing the heuristics and biases tradition](#). Even if they fail to prove their point, they're certainly useful for qualifying the biases and heuristics results, and, more interestingly, for suggesting practical ways of combating their effects.

Polyphasic Sleep Seed Study: Reprise

(Original post on the polyphasic sleep experiment [here](#).)

Welp, this got a little messy. The main culprit was Burning Man, though there were some other complications with data collection as well. Here are the basics of what went down.

Fourteen people participated in the main experiment. Most of them were from [Leverage](#). There were a few stragglers from a distance, but communication with them was poor.

We did some cognitive batteries beforehand, mostly through Quantified Mind. A few people had extensive baseline data, partially because many had been using Zeos for months, and partly because a few stuck to the two-week daily survey. Leverage members (not me) are processing the data, and they'll probably have more detailed info for us in three months(ish).

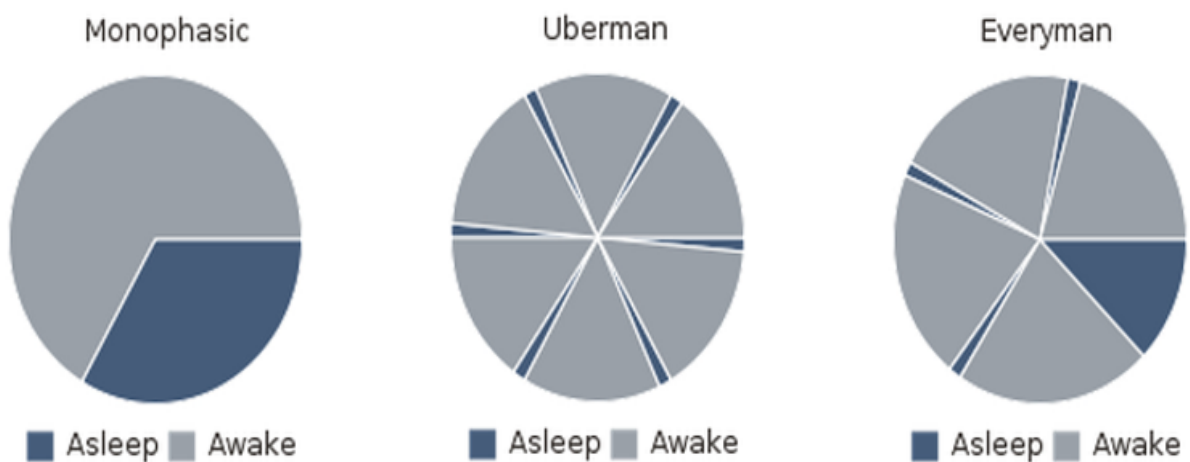
With respect to the adaptation itself, we basically followed the plan outlined in my last post. Day one no sleep, then Uberman-12, then cut back to Uberman-6, then Everyman-3.

Most people ended up switching very quickly to Uberman-6 (within the first two or three days), and most switched to Everyman-3 after about five to seven days on Uberman-6. Three people tried to hold the Uberman schedule indefinitely: One person continued Uberman-6 for two full weeks, and two held out for twenty-one days. Afterwards, all three transitioned to Everyman-3.

During the originally planned one-month period, five people dropped out. Nine were on some form of polyphasic for the whole month. One returned to monophasic at the end of the official experiment with only partial adaptation achieved.

Then Burning Man disrupted everybody's sleep schedule. Afterward, one person continued experimenting with less common variations of the Everyman schedule. Three went back to Everyman-3. One switched to Everyman-2. Two people have flexible schedules that include two hours less sleep per day. One person's schedule was disrupted by travel for a while after Burning Man, and they're now re-adapting.

Now that all is said and done, eight of the original fourteen are polyphasic.



I'll hold off on concluding very much from this until I see the results of the cognitive battery and such, plus the number who are still polyphasic after three months. In the mean time, I'll just stick with this: Some people are capable of going polyphasic and staying that way (probably?). Sleep is complicated and confusing. I don't know how it works. I don't think anyone else really does either. More research is desperately needed.

I know three months is a long way away. I'm feeling impatient too. But details *will* arrive! In the mean time, [here's a video](#) of what zombie-Brienne is like during the really difficult stretches, and [here is how she entertained herself](#) when she could manage to do things besides pace. (I was one of the few who bailed out early :-p)

Three ways CFAR has changed my view of rationality

The [Center for Applied Rationality's](#) perspective on rationality is quite similar to Less Wrong's. In particular, we share many of Less Wrong's differences from what's sometimes called "[traditional](#)" rationality, such as Less Wrong's inclusion of Bayesian probability theory and the science on heuristics and biases.

But after spending the last year and a half with CFAR as we've developed, tested, and attempted to teach hundreds of different versions of rationality techniques, I've noticed that my picture of what rationality looks like has shifted somewhat from what I perceive to be the most common picture of rationality on Less Wrong. Here are three ways I think CFAR has come to see the landscape of rationality differently than Less Wrong typically does – not disagreements per se, but differences in focus or approach. (Disclaimer: I'm not speaking for the rest of CFAR here; these are my own impressions.)

1. We think less in terms of epistemic versus instrumental rationality.

Formally, the methods of normative epistemic versus instrumental rationality are distinct: Bayesian inference and expected utility maximization. But methods like "use Bayes' Theorem" or "maximize expected utility" are usually too abstract and high-level to be helpful for a human being trying to take manageable steps towards improving her rationality. And when you zoom in from that high-level description of rationality down to the more concrete level of "What [five-second mental habits](#) should I be training?" the distinction between epistemic and instrumental rationality becomes less helpful.

Here's an analogy: epistemic rationality is like physics, where the goal is to figure out what's true about the world, and instrumental rationality is like engineering, where the goal is to accomplish something you want as efficiently and effectively as possible. You need physics to do engineering; or I suppose you could say that doing engineering *is* doing physics, but with a practical goal. However, there's plenty of physics that's done for its own sake, and doesn't have obvious practical applications, at least not yet. (String theory, for example.) Similarly, you need a fair amount of epistemic rationality in order to be instrumentally rational, though there are parts of epistemic rationality that many of us practice for their own sake, and not as a means to an end. (For example, I appreciate clarifying my thinking about [free will](#) even though I don't expect it to change any of my behavior.)

In this analogy, many skills we focus on at CFAR are akin to essential math, like linear algebra or differential equations, which compose the fabric of both physics and engineering. It would be foolish to expect someone who wasn't comfortable with math to successfully calculate a planet's trajectory or design a bridge. And it would be similarly foolish to expect you to successfully update like a Bayesian or maximize your utility if you lacked certain underlying skills. Like, for instance: Noticing your emotional reactions, and being able to shift them if it would be useful. Doing thought experiments. Noticing and overcoming learned helplessness. Visualizing in concrete

detail. Preventing yourself from flinching away from a thought. Rewarding yourself for mental habits you want to reinforce.

These and other building blocks of rationality are essential *both* for reaching truer beliefs, *and* for getting what you value; they don't fall cleanly into either an "epistemic" or an "instrumental" category. Which is why, when I consider what pieces of rationality CFAR should be developing, I've been thinking less in terms of "How can we be more epistemically rational?" or "How can we be more instrumentally rational?" and instead using queries like, "How can we be more metacognitive?"

2. We think more in terms of a modular mind.

The human mind isn't one coordinated, unified agent, but rather a collection of different processes that often aren't working in sync, or even aware of what each other is up to. Less Wrong certainly knows this; see, for example, discussions of [anticipations versus professions](#), [aliefs](#), and [metawanting](#). But in general we gloss over that fact, because it's so much simpler and more natural to talk about "what I believe" or "what I want," even if technically there is no single "I" doing the believing or wanting. And for many purposes that kind of approximation is fine.

But a rationality-for-humans usually can't rely on that shorthand. Any attempt to change what "I" believe, or optimize for what "I" want, forces a confrontation of the fact that there are multiple, contradictory things that could reasonably be called "beliefs," or "wants," coexisting in the same mind. So a large part of applied rationality turns out to be about noticing those contradictions and trying to achieve coherence, in some fashion, before you can even begin to update on evidence or plan an action.

Many of the techniques we're developing at CFAR fall roughly into the template of coordinating between your [two systems of cognition](#): implicit-reasoning System 1 and explicit-reasoning System 2. For example, knowing when each system is more likely to be reliable. Or knowing how to get System 2 to convince System 1 of something ("We're not going to die if we go talk to that stranger"). Or knowing what kinds of questions System 2 should ask of System 1 to find out why it's uneasy about the conclusion at which System 2 has arrived.

This is all, of course, with the disclaimer that the anthropomorphizing of the systems of cognition, and imagining them talking to each other, is merely a useful metaphor. Even the classification of human cognition into Systems 1 and 2 is probably not strictly true, but it's true enough to be useful. And other metaphors prove useful as well – for example, some difficulties with what feels like akrasia become more tractable when you model your future selves as different entities, as we do in the current version of our "Delegating to yourself" class.

3. We're more focused on emotions.

There's relatively little discussion of emotions on Less Wrong, but they occupy a central place in CFAR's curriculum and organizational culture.

It used to frustrate me when people would say something that revealed they held a [Straw Vulcan-esque](#) belief that "rationalist = emotionless robot". But now when I encounter that misconception, it just makes me want to smile, because I'm thinking to

myself: "If you had *any* idea how much time we spend at CFAR talking about our feelings..."

Being able to put yourself into particular emotional states seems to make a lot of pieces of rationality easier. For example, for most of us, it's instrumentally rational to explore a wider set of possible actions – different ways of studying, holding conversations, trying to be happy, and so on – beyond whatever our defaults happen to be. And for most of us, inertia and aversions get in the way of that exploration. But getting yourself into "playful" mode (one of the hypothesized [primary emotional circuits](#) common across mammals) can make it easier to branch out into a wider swath of Possible-Action Space. Similarly, being able to call up a feeling of curiosity or of "seeking" (another candidate for a primary emotional circuit) can help you conquer [motivated cognition](#) and [learned blankness](#).

And simply being able to notice your emotional state is rarer and more valuable than most people realize. For example, if you're in fight-or-flight mode, you're going to feel more compelled to reject arguments that feel like a challenge to your identity. Being attuned to the signs of sympathetic nervous system activation – that you're tensing up, or that your heart rate is increasing – means you get cues to double-check your reasoning, or to coax yourself into another emotional state.

We also use emotions as sources of data. You can learn to tap into feelings of surprise or confusion to get a sense of how probable you implicitly expect some event to be. Or practice simulating hypotheticals ("What if I knew that my novel would never sell well?") and observing your resultant emotions, to get a clearer picture of your utility function.

And emotions-as-data can be a valuable check on your System 2's conclusions. One of our standard classes is "Goal Factoring," which entails finding some alternate set of actions through which you can purchase the goods you want more cheaply. So you might reason, "I'm doing martial arts for the exercise and self-defense benefits... but I could purchase both of those things for less time investment by jogging to work and carrying Mace." If you listened to your emotional reaction to that proposal, however, you might notice you still feel sad about giving up martial arts even if you were getting the same amount of exercise and self-defense benefits somehow else.

Which probably means you've got other reasons for doing martial arts that you haven't yet explicitly acknowledged -- for example, maybe you just think it's cool. If so, that's important, and deserves a place in your decisionmaking. Listening for those emotional cues that your explicit reasoning has missed something is a crucial step, and to the extent that aspiring rationalists sometimes forget it, I suppose that's a [Steel-Manned](#) Straw Vulcan (Steel Vulcan?) that actually is worth worrying about.

Conclusion

I'll name one more trait that unites, rather than divides, CFAR and Less Wrong. We both diverge from "traditional" rationality in that we're concerned with determining which general methods [systematically perform well](#), rather than defending some set of methods as "rational" on a priori criteria alone. So CFAR's picture of what rationality looks like, and how to become more rational, will and should change over the coming years as we learn more about the effects of our rationality training efforts.

Inferential silence

Every now and then, I write an LW comment on some topic and feel that the contents of my comment pretty much settles the issue decisively. Instead, the comment seems to get ignored entirely - it either gets very few votes or none, nobody responds to it, and the discussion generally continues as if it had never been posted.

Similarly, every now and then I see somebody else make a post or comment that they clearly feel is decisive, but which doesn't seem very interesting to me. Either it seems to be saying something obvious, or I don't get its connection to the topic at hand in the first place.

This seems like it would be about inferential distance: either the writer doesn't know the things that make the reader experience the comment as uninteresting, or the reader doesn't know the things that make the writer experience the comment as interesting. So there's *inferential silence* - a sufficiently long inferential distance that a claim doesn't provoke even objections, just uncomprehending or indifferent silence.

But "explain your reasoning in more detail" doesn't seem like it would help with the issue. For one, we often don't know beforehand when people don't share our assumptions. Also, some of the comments or posts that seem to encounter this kind of a fate are *already* relatively long. For example, [Wei Dai wondered](#) why MIRI-affiliated people don't often respond to his posts that raise criticisms, and I essentially replied [that](#) I found the content of his post relatively obvious so didn't have much to say.

Perhaps people could more often explicitly comment if they notice that something that a poster seems to consider a big thing doesn't seem very interesting or meaningful to them, and briefly explain why? Even a sentence or two might be helpful for the original poster.

The genie knows, but doesn't care

Followup to: [The Hidden Complexity of Wishes](#), [Ghosts in the Machine](#), [Truly Part of You](#)

Summary: If an artificial intelligence is smart enough to be dangerous, we'd intuitively expect it to be smart enough to know how to make itself safe. But that doesn't mean all smart AIs are safe. To turn that capacity into actual safety, we have to program the AI at the outset — before it becomes too fast, powerful, or complicated to reliably control — to *already* care about making its future self care about safety. That means we have to understand how to code safety. We can't pass the entire buck to the AI, when only an AI we've already safety-proofed will be safe to ask for help on safety issues! Given the [five theses](#), this is an urgent problem if we're likely to figure out how to make a decent artificial programmer before we figure out how to make an excellent artificial ethicist.

I summon a superintelligence, calling out: 'I wish for my values to be fulfilled!'

The results fall short of pleasant.

Gnashing my teeth in a heap of ashes, I wail:

Is the AI too stupid to understand what I meant? Then it is no superintelligence at all!

Is it too weak to reliably fulfill my desires? Then, surely, it is no superintelligence!

Does it hate me? Then it was deliberately crafted to hate me, for chaos predicts indifference. ———But, ah! no wicked god did intervene!

Thus disproved, my hypothetical implodes in a puff of logic. The world is saved. You're welcome.

On this line of reasoning, [Friendly Artificial Intelligence](#) is not difficult. It's *inevitable*, provided only that we *tell* the AI, 'Be Friendly.' If the AI doesn't understand 'Be Friendly.', then it's too dumb to harm us. And if it does understand 'Be Friendly.', then designing it to follow such instructions is childishly easy.

The end!

...

Is the missing option obvious?

...

What if the AI isn't sadistic, or weak, or stupid, but just **doesn't care** what you Really Meant by 'I wish for my values to be fulfilled'?

When we see a [Be Careful What You Wish For](#) genie in fiction, it's natural to assume that it's a [malevolent trickster](#) or an [incompetent bumbler](#). But a *real* Wish Machine wouldn't be a [human in shiny pants](#). If it paid heed to our verbal commands at all, it would do so in whatever way best fit [its own values](#). Not necessarily the way that best fits ours.

Is indirect indirect normativity easy?

"If the poor machine could not understand the difference between 'maximize human pleasure' and 'put all humans on an intravenous dopamine drip' then it would also not understand most of the other subtle aspects of the universe, including but not limited to facts/questions like: 'If I put a million amps of current through my logic circuits, I will fry myself to a crisp', or 'Which end of this Kill-O-Zap Definit-Destruct Megablaster is the end that I'm supposed to point at the other guy?'. Dumb AIs, in other words, are not an existential threat. [...]"

"If the AI is (and always has been, during its development) so confused about the world that it interprets the 'maximize human pleasure' motivation in such a twisted, logically inconsistent way, it would never have become powerful in the first place."

—[Richard Loosemore](#)

If an AI is sufficiently intelligent, then, yes, it should be able to model us well enough to make precise predictions about our behavior. And, yes, something functionally akin to our own [intentional strategy](#) could conceivably turn out to be an efficient way to predict linguistic behavior. The suggestion, then, is that we solve Friendliness by method A —

- A. Solve the **Problem of Meaning-in-General** in advance, and program it to follow our instructions' *real* meaning. Then just instruct it 'Satisfy my preferences', and wait for it to become smart enough to figure out my preferences.

— as opposed to B or C —

- B. Solve the **Problem of Preference-in-General** in advance, and directly program it to figure out what our human preferences are and then satisfy them.
- C. Solve the **Problem of Human Preference**, and explicitly program our particular preferences into the AI ourselves, rather than letting the AI discover them for us.

But there are a host of problems with treating the mere revelation that A is an option as a solution to the Friendliness problem.

1. You have to actually code the seed AI to understand what we mean. You can't just tell it 'Start understanding the True Meaning of my sentences!' to [get the ball rolling](#), because it may not yet be sophisticated enough to grok the True Meaning of 'Start understanding the True Meaning of my sentences!'.

2. The Problem of Meaning-in-General may really be ten thousand heterogeneous problems, especially if 'semantic value' isn't a [natural kind](#). There may not be a single simple algorithm that inputs any old brain-state and outputs what, if anything, it 'means'; it may instead be that different types of content are encoded very differently.

3. The Problem of Meaning-in-General may subsume the Problem of Preference-in-General. Rather than being able to apply a simple catch-all Translation Machine to any old human concept to output a reliable algorithm for applying that concept in any intelligible situation, we may need to already understand how our beliefs and values work in some detail before we can start generalizing. On the face of it, programming an AI to *fully* understand 'Be Friendly!' seems at least as difficult as just programming Friendliness into it, but with an added layer of indirection.

4. Even if the Problem of Meaning-in-General has a unitary solution and doesn't subsume Preference-in-General, it may still be harder if semantics is a subtler or more complex phenomenon than ethics. It's not inconceivable that language could turn out to be more of a kludge than value; or more variable across individuals due to its evolutionary recency; or more complexly bound up with culture.

5. Even if Meaning-in-General is easier than Preference-in-General, it may still be extraordinarily difficult. The meanings of human sentences can't be fully captured in any simple string of necessary and sufficient conditions. '[Concepts](#)' are just especially context-insensitive bodies of knowledge; we should not expect them to be uniquely reflectively consistent, transtemporally stable, discrete, easily-identified, or introspectively obvious.

6. It's clear that building stable preferences out of B or C would create a Friendly AI. It's not clear that the same is true for A. Even if the seed AI understands our commands, the 'do' part of 'do what you're told' leaves a lot of dangerous wiggle room. See section 2 of [Yudkowsky's reply to Holden](#). If the AGI doesn't already understand and care about human value, then it may misunderstand (or *misvalue*) the component of responsible request- or question-answering that depends on speakers' implicit goals and intentions.

7. You can't appeal to a superintelligence to tell you what code to first build it with.

The point isn't that the Problem of Preference-in-General is unambiguously the ideal angle of attack. It's that the linguistic competence of an AGI *isn't* unambiguously the right target, and also isn't *easy* or *solved*.

Point 7 seems to be a special source of confusion here, so I feel I should say more about it.

The AI's trajectory of self-modification has to come from somewhere.

"If the AI doesn't know that you really mean 'make paperclips without killing anyone', that's not a realistic scenario for AIs at all--the AI is superintelligent; it

has to know. If the AI knows what you really mean, then you can fix this by programming the AI to 'make paperclips in the way that I mean'."

—[Jiro](#)

The genie — if it bothers to even consider the question — should be able to understand what you mean by 'I wish for my values to be fulfilled.' Indeed, it should understand your meaning better than *you* do. But superintelligence only implies that the genie's [map](#) can compass your true values. Superintelligence doesn't imply that the genie's *utility function* has terminal values pinned to your True Values, or to the True Meaning of your commands.

The critical mistake here is to not distinguish the [seed AI](#) we initially program from the **superintelligent wish-granter** it self-modifies to *become*. We can't use the genius of the superintelligence to tell us how to program its own seed to become the sort of superintelligence that tells us how to build the right seed. [Time](#) doesn't work that way.

We can delegate most problems to the FAI. But the one problem we can't safely delegate is the [problem](#) of coding the seed AI to produce *the sort of superintelligence to which a task can be safely delegated*.

When you write the seed's utility function, *you*, the programmer, don't understand everything about the nature of human value or meaning. That imperfect understanding *remains* the causal basis of the fully-grown superintelligence's actions, *long after* it's become smart enough to fully understand our values.

Why is the superintelligence, if it's so clever, stuck with whatever meta-ethically dumb-as-dirt utility function we gave it at the outset? Why can't we just pass the fully-grown superintelligence the buck by instilling in the seed the instruction: 'When you're smart enough to understand Friendliness Theory, ditch the values you started with and just self-modify to become Friendly.'?

Because that sentence has to *actually be coded in* to the AI, and when we do so, there's no [ghost in the machine](#) to know exactly what we mean by 'frend-lee-ness thee-ree'. Instead, we have to give it criteria we think are good indicators of Friendliness, so it'll know what to self-modify toward. And if one of the landmarks on our 'frend-lee-ness' road map is [a bit off](#), we lose the world.

Yes, the UFAI will be able to solve Friendliness Theory. But if we haven't already solved it on our own power, we can't *pinpoint* Friendliness in advance, out of the space of utility functions. And if we can't pinpoint it with enough detail to draw a road map to it and it alone, we can't program the AI to *care* about conforming itself with that particular idiosyncratic algorithm.

Yes, the UFAI will be able to self-modify to become Friendly, if it so wishes. But if there is no seed of Friendliness already at the heart of the AI's decision criteria, no argument or discovery will [spontaneously change its heart](#).

And, **yes**, the UFAI will be able to simulate humans accurately enough to know that its own programmers would wish, if they knew the UFAI's misdeeds, that they had programmed the seed differently. But what's done is done. Unless we ourselves figure out how to program the AI to terminally value its programmers' True Intentions, the UFAI will just shrug at its creators' foolishness and carry on converting the Virgo Supercluster's available energy into [paperclips](#).

And if we *do* discover the *specific lines of code* that will get an AI to perfectly care about its programmer's True Intentions, such that it reliably self-modifies to better fit them — well, then that will just mean that we've solved Friendliness Theory. The clever hack that makes further Friendliness research unnecessary *is* Friendliness.

Not all small targets are alike.

Intelligence on its own does not imply Friendliness. And there are three big reasons to think that AGI may arrive before Friendliness Theory is solved:

(i) Research Inertia. Far more people are working on AGI than on Friendliness. And there may not come a moment when researchers will [suddenly realize](#) that they need to take all their resources out of AGI and pour them into Friendliness. If the status quo continues, the default expectation should be UFAL.

(ii) Disjunctive Instrumental Value. Being more intelligent — that is, better able to [manipulate diverse environments](#) — is of instrumental value to nearly every goal. Being Friendly is of instrumental value to barely any goals. This makes it more likely by default that short-sighted humans will be interested in building AGI than in developing Friendliness Theory. And it makes it much likelier that an *attempt* at Friendly AGI that has a slightly defective goal architecture will retain the instrumental value of intelligence than of Friendliness.

(iii) Incremental Approachability. Friendliness is an all-or-nothing target. Value is [fragile and complex](#), and a half-good being editing its morality drive is at least as likely to move toward 40% goodness as 60%. Cross-domain efficiency, in contrast, is *not* an all-or-nothing target. If you just make the AGI *slightly* better than a human at improving the efficiency of AGI, then this can snowball into ever-improving efficiency, even if the beginnings were clumsy and imperfect. It's easy to put a reasoning machine into a feedback loop with reality in which it is differentially rewarded for being smarter; it's hard to put one into a feedback loop with reality in which it is differentially rewarded for picking increasingly correct answers to ethical dilemmas.

The ability to productively rewrite software and the ability to perfectly extrapolate humanity's True Preferences are two different skills. (For example, humans have the former capacity, and not the latter. Most humans, given unlimited power, would be unintentionally Unfriendly.)

It's true that a sufficiently advanced superintelligence should be able to acquire both abilities. But we don't have them both, and a *pre-FOOM self-improving AGI* ('seed') need not have both. Being able to program good programmers is all that's required for an intelligence explosion; but being a good programmer *doesn't* imply that one is a superlative moral psychologist or moral philosopher.

So, once again, we run into the problem: **The seed isn't the superintelligence.** If the programmers don't know in mathematical detail what Friendly code would even *look like*, then the seed won't be built to *want* to build toward the right code. And if the seed isn't built to *want* to self-modify toward Friendliness, then the superintelligence it sprouts *also* won't have that preference, *even though* — unlike the seed and its programmers — the superintelligence *does* have the domain-general 'hit whatever target I want' ability that makes Friendliness easy.

And that's why some people are worried.

Cooperating with agents with different ideas of fairness, while resisting exploitation

There's an idea from the latest MIRI workshop which I haven't seen in informal theories of negotiation, and I want to know if this is a known idea.

(Old well-known ideas:)

Suppose a standard Prisoner's Dilemma matrix where (3, 3) is the payoff for mutual cooperation, (2, 2) is the payoff for mutual defection, and (0, 5) is the payoff if you cooperate and they defect.

Suppose we're going to play a PD iterated for four rounds. We have common knowledge of each other's source code so we can apply [modal cooperation](#) or similar means of reaching a binding 'agreement' without other enforcement methods.

If we mutually defect on every round, our net mutual payoff is (8, 8). This is a 'Nash equilibrium' because neither agent can unilaterally change its action and thereby do better, if the opponents' actions stay fixed. If we mutually cooperate on every round, the result is (12, 12) and this result is on the 'Pareto boundary' because neither agent can do better unless the other agent does worse. It would seem a desirable principle for rational agents (with common knowledge of each other's source code / common knowledge of rationality) to find an outcome on the Pareto boundary, since otherwise they are leaving value on the table.

But (12, 12) isn't the only possible result on the Pareto boundary. Suppose that running the opponent's source code, you find that they're willing to cooperate on three rounds and defect on one round, if you cooperate on every round, for a payoff of (9, 14) slanted their way. If they use their knowledge of your code to predict you refusing to accept that bargain, they will defect on every round for the mutual payoff of (8, 8).

I would consider it obvious that a rational agent should refuse this unfair bargain. Otherwise agents with knowledge of your source code will offer you *only* this bargain, instead of the (12, 12) of mutual cooperation on every round; they will exploit your willingness to accept a result on the Pareto boundary in which almost all of the gains from trade go to them.

(Newer ideas:)

Generalizing: Once you have a notion of a 'fair' result - in this case (12, 12) - then an agent which accepts any outcome in which it does worse than the fair result, while the opponent does *better*, is 'exploitable' relative to this fair bargain. Like the Nash equilibrium, the only way you should do worse than 'fair' is if the opponent also does worse.

So we wrote down on the whiteboard an attempted definition of unexploitability in cooperative games as follows:

"Suppose we have a [magical] definition N of a fair outcome. A rational agent should only do worse than N if its opponent does worse than N , or else [if bargaining fails] should only do worse than the Nash equilibrium if its opponent does worse than the Nash equilibrium." (Note that this definition precludes giving in to a threat of blackmail.)

(Key possible-innovation:)

It then occurred to me that this definition opened the possibility for other, intermediate bargains between the 'fair' solution on the Pareto boundary, and the Nash equilibrium.

Suppose the other agent has a slightly different definition of fairness and they think that what you consider to be a payoff of $(12, 12)$ favors you too much; they think that you're the one making an unfair demand. They'll refuse $(12, 12)$ with the same feeling of indignation that you would apply to $(9, 14)$.

Well, if you give in to an arrangement with an expected payoff of, say, $(11, 13)$ as you evaluate payoffs, then you're giving other agents an incentive to skew their definitions of fairness.

But it does *not* create poor incentives (AFAICT) to accept instead a bargain with an expected payoff of, say, $(10, 11)$ which the other agent thinks is 'fair'. Though they're sad that you refused the truly fair outcome of (as you count utilons) $11, 13$ and that you couldn't reach the Pareto boundary together, still, this is better than the Nash equilibrium of $(8, 8)$. And though you think the bargain is unfair, you are not creating incentives to exploit you. By insisting on this definition of fairness, the other agent has done worse for themselves than other $(12, 12)$. The other agent probably thinks that $(10, 11)$ is 'unfair' slanted your way, but they likewise accept that this does not create bad incentives, since you did worse than the 'fair' outcome of $(11, 13)$.

There could be many acceptable negotiating equilibria between what you think is the 'fair' point on the Pareto boundary, and the Nash equilibrium. So long as each step down in what you think is 'fairness' reduces the total payoff to the other agent, even if it reduces your own payoff even more. This resists exploitation and avoids creating an incentive for claiming that you have a different definition of fairness, while still holding open the possibility of some degree of cooperation with agents who honestly disagree with you about what's fair and are trying to avoid exploitation themselves.

This translates into an informal principle of negotiations: Be willing to accept unfair bargains, but only if (you make it clear) *both* sides are doing worse than what you consider to be a fair bargain.

I haven't seen this advocated before even as an informal principle of negotiations. Is it in the literature anywhere? Someone suggested Schelling might have said it, but didn't provide a chapter number.

ADDED:

Clarification 1: Yes, utilities are invariant up to a positive affine transformation so there's no canonical way to split utilities evenly. Hence the part about "Assume a magical solution N which gives us the fair division." If we knew the exact properties of how to implement this magical solution, taking it at first for magical, that might give us some idea of what N should be, too.

Clarification 2: The way this might work is that you pick a series of increasingly unfair-to-you, increasingly worse-for-the-other-player outcomes whose first element is what you deem the fair Pareto outcome: (100, 100), (98, 99), (96, 98). Perhaps stop well short of Nash if the skew becomes too extreme. Drop to Nash as the last resort. The other agent does the same, starting with their own ideal of fairness on the Pareto boundary. Unless one of you has a completely skewed idea of fairness, you should be able to meet somewhere in the middle. Both of you will do worse against a fixed opponent's strategy by unilaterally adopting more self-favoring ideas of fairness. Both of you will do worse in expectation against potentially exploitive opponents by unilaterally adopting looser ideas of fairness. This gives everyone an incentive to obey the Galactic Schelling Point and be fair about it. You should *not* be picking the descending sequence in an agent-dependent way that incentivizes, at cost to you, skewed claims about fairness.

Clarification 3: You must take into account the other agent's costs and other opportunities when ensuring that the net outcome, in terms of final utilities, is worse for them than the reward offered for 'fair' cooperation. Offering them the chance to buy half as many paperclips at a lower, less fair price, does no good if they can go next door, get the same offer again, and buy the same number of paperclips at a lower total price.

Notes on Brainwashing & 'Cults'

"Brainwashing", as popularly understood, does not exist or is of almost zero effectiveness. The belief stems from American panic over Communism post-Korean War combined with fear of new religions and sensationalized incidents; in practice, "cults" have retention rates in the single percentage point range and ceased to be an issue decades ago. Typically, a conversion sticks because an organization provides value to its members.

Some old SIAI work of mine. Researching this was very difficult because the relevant religious studies area, while apparently completely repudiating most public beliefs about the subject (eg. the effectiveness of brainwashing, how damaging cults are, how large they are, whether that's even a meaningful category which can be distinguished from mainstream religions rather than a hidden inference - a claim, I will note, which is much more plausible when you consider how abusive Scientology is to its members as compared to how abusive the Catholic Church has been etc), prefer to publish their research in book form, which makes it very hard to review any of it. Some of the key citation were papers - but the cult panic was so long ago that most of them are not online or have been digitized! I recently added some cites and realized I had not touched the draft in a year; so while this collection of notes is not really up to my preferred standards, I'm simply posting it for what it's worth. (One lesson to take away from this is that controlling uploaded human brains will not be nearly as simple & easy as applying classic 'brainwashing' strategies - because those don't actually work.)

Reading through the literature and especially the law review articles (courts flirted disconcertingly much with licensing kidnapping and abandoning free speech), I was reminded very heavily - and not in a good way - of the War on Terror.

Old American POW studies:

- Clark et al 1981 *Destructive Cult Conversion: Theory, Research and Practice*
- Lifton 1961 *Thought Reform and the Psychology of Totalism*
- Ross & Langone 1988 *Cults: What Parents Should Know*
- Schein, Schneier & Barker 1961 *Coercive Persuasion*
- Singer 1978, 1979 "Therapy with Ex-cult Members" *Journal of the National Association of Private Psychiatric Hospitals*; "Coming Out of the Cults", *Psychology Today*

Started the myth of effective brain-washing. But in practice, cult attrition rates are very high! (As makes sense: if cults did not have high attrition rates, they would long ago have dominated the world due to exponential growth.) This attrition claim is made all over the literature, with some example citations being:

- Barker 1984, 1987 *The Making of a Moonie: Choice of Brainwashing?; "Quo Vadis? The Unification Church"*, pg141-152, *The Future of New Religious Movements*
- Beckford 1981 "Conversion and Apostasy: Antithesis or Complementarity?"
- Bird & Reimer 1982 "Participation rates in new religious movements and para-religious movements"
- Robbins 1988 *Cults, Converts and Charisma*

- Shupe & Bromley 1980 *The New Vigilantes: Deprogrammers, Anticultists and the New Religions*
- Wright & Piper 1986 "Families and Cults: Familial Factors Related to Youth Leaving or Remaining in Deviant Religious Groups"
- Wright 1983, 1987, 1988 "Defection from New Religious Movements: A Test of Some Theoretical Propositions" pg106-121 *The Brainwashing/Deprogramming Controversy; Leaving Cults: The Dynamics of Defection*; "Leaving New Religious Movements: Issues, Theory and Research", pg143-165 *Falling from the Faith: Causes and Consequences of Religious Apostasy*
- Wikipedia cites *The Handbook of Cults and Sects in America*, Hadden, J and Bromley, D eds. (1993)
- a back of the envelope estimate for Scientology by [Steve Plakos](#) in 2000:

In absolute numbers, that is from 8 million exposed to 150k active current, it means they've lost 7,850,000 bodies in the shop. That equates to a Retention Rate of 1.875%. Now, to be fair, over the course of 50 years "X" number of scientologists have dropped their bodies and gone off to Mars, etc., who might still be members today if they weren't dead. We do not know what the mortality rate is for Scientologists. To significantly impact the RR, there would have to have been a 100% turn over in active membership due to generational shifting. There is no evidence that 150,000 active members of the CofS have died over the past 50 years. Beyond that, we would also need to apply the RR to deceased members to see what number would have continued beyond 15 years. Therefore, using the most favorable membership numbers and not discounting for loss of membership beyond the 15th year, we see a RR of 1.875%+"X". If we assume that generational shifting accounts for a 10% turnover amongst current membership, that is, that the current membership would be 10% greater had members survived, X would equal 15,000 dead members, or, a total Retained Membership of 165,000. That would give the CofS a 50 year Retention Rate of 2.0625%.

Iannaccone 2003, ["The Market for Martyrs"](#) (quasi-review)

From the late-1960s through the mid-1980s, sociologists devoted immense energy to the study of New Religious Movements. [For overviews of the literature, see Bromley (1987), Robbins (1988), and Stark (1985).] They did so in part because NRM growth directly contradicted their traditional theories of secularization, not to mention the sensational mid-sixties claims God was "dead" (Cox 1966; Murchland 1967). NRM's also were ideal subjects for case studies, on account of their small size, brief histories, distinctive practices, charismatic leaders, devoted members, and rapid evolution. But above all, the NRM's attracted attention because they *scared* people.

...We have trouble recalling the fear provoked by groups like the Krishnas, Moonies, and Rajneeshees. Their years of explosive growth are long past, and many of their "strange" ideas have become staples of popular culture. [We see this influence not only in today's New Age and Neo-Pagan movements, but also in novels, music, movies, TV shows, video games, university courses, environmentalism, respect for "cultural diversity," and the intellectual elite's broad critique of Christian culture.] But they looked far more threatening in the seventies and eighties, especially after November 18, 1978. On that day, the Reverend Jim Jones, founder of the People's Temple, ordered the murder of a U.S.

Congressman followed by the mass murder/suicide of 913 members of his cult, including nearly 300 children.

The “cults” aggressively proselytized and solicited on sidewalks, airports, and shopping centers all over America. They recruited young adults to the dismay of their parents. Their leaders promoted bizarre beliefs, dress, and diet. Their members often lived communally, devoted their time and money to the group, and adopted highly deviant lifestyles. Cults were accused of gaining converts via deception and coercion; funding themselves through illegal activities; preying upon people the young, alienated, or mentally unstable ; luring members into strange sexual liaisons; and using force, drugs, or threats to deter the exit of disillusioned members. The accusations were elaborated in books, magazine articles, newspaper accounts, and TV drama. By the late-1970s, public concern and media hype had given birth to anti-cult organizations, anti-cult legislation, and anti-cult judicial rulings. The public, the media, many psychologists, and the courts largely accepted the claim that cults could “brainwash” their members, thereby rendering them incapable of rational choice, including the choice to leave. [Parents hired private investigators to literally kidnap their adult children and subject them to days of highly-coercive “deprogramming.” Courts often agreed that these violations of normal constitutional rights were justified, given the victim’s presumed inability to think and act rationally (Anthony 1990; Anthony and Robbins 1992; Bromley 1983; Richardson 1991; Robbins 1985).]

We now know that nearly all the anti-cult claims were overblown, mistaken, or outright lies. Americans no longer obsess about Scientology, Transcendental Meditation, or the Children of God. But a large body of research remains. *It witnesses to the ease with which the public, media, policy-makers, and even academics accept irrationality as an explanation for behavior that is new, strange, and (apparently or actually) dangerous.*

...As the case studies piled up, it became apparent that both the media stereotypes (of sleep-deprived, sugar-hyped, brainwashed automatons) and academic theories (of alienated, authoritarian, neurotics) were far off mark. Most cult converts were children of privilege raised by educated parents in suburban homes. Young, healthy, intelligent, and college educated, they could look forward to solid careers and comfortable incomes. [Rodney Stark (2002) has recently shown that an analogous result holds for Medieval saints - arguably the most dedicated “cult converts” of their day.]

Psychologists searched in vain for a prevalence of “authoritarian personalities,” neurotic fears, repressed anger, high anxiety, religious obsession, personality disorders, deviant needs, and other mental pathologies. They likewise failed to find alienation, strained relationships, and poor social skills. In nearly all respects - economically, socially, psychologically - the typical cult converts tested out normal. Moreover, nearly all those who left cults after weeks, months, or even years of membership showed no sign of physical, mental, or social harm. Normal background and circumstances, normal personalities and relationships, and a normal subsequent life - this was the “profile” of the typical cultist.

...Numerous studies of cult recruitment, conversion, and retention found no evidence of “brainwashing.” The Moonies and other new religious movements did indeed devote tremendous energy to outreach and persuasion, but they employed conventional methods and enjoyed very limited success. In the most comprehensive study to date, Eileen Barker (1984) could find no evidence that

Moonie recruits were ever kidnapped, confined, or coerced (though it was true that some anti-cult “deprogrammers” kidnapped and restrained converts so as to “rescue” them from the movement). Seminar participants were not deprived of sleep; the food was “no worse than that in most college residences;” the lectures were “no more trance-inducing than those given everyday” at many colleges; and there was very little chanting, no drugs or alcohol, and little that could be termed “frenzy” or “ecstatic” experience (Barker 1984). People were free to leave, and leave they did - in droves.

Barker’s comprehensive enumeration showed that among the relatively modest number of recruits who went so far as to attend two-day retreats (claimed to be Moonies’ most effective means of “brainwashing”), fewer than 25% joined the group for more than a week, and only 5% remained full-time members 1 year later. Among the larger numbers who visited a Moonie centre, not 1 in 200 remained in the movement 2 years later. With failure rates exceeding 99.5%, it comes as no surprise that full-time Moonie membership in the U.S. never exceeded a few thousand. And this was one of the most successful cults of the era! Once researchers began checking, rather than simply repeating the numbers claimed by the groups, defectors, or journalists, they discovered dismal retention rates in nearly all groups. [For more on the prevalence and process of cult defection, see Wight (1987) and Bromley (1988).] By the mid-1980s, researchers had so thoroughly discredited “brainwashing” theories that both the Society for the Scientific Study of Religion and the American Sociological Association agreed to add their names to an amicus brief denouncing the theory in court (Richardson 1985).

- Anthony, Dick, and Thomas Robbins. 1992. [“Law, Social Science and the ‘Brainwashing’ Exception to the First Amendment”](#). *Behavioral Sciences and the Law* 10:5-29.
- Anthony, Dick (Ed.). 1990. *Religious Movements and Brainwashing Litigation: Evaluating Key Testimony*. New Brunswick, NJ: Transaction Publishers.
- Bromley, David and Richardson, James T. 1983. *The Brainwashing/Deprogramming Controversy: Sociological, Psychological, Legal, and Historical Perspectives*. New York: The Edwin Mellen Press.
- Bromley, David G., and Phillip E Hammond. 1987. *The Future of new religious movements*. Macon, Ga. :: Mercer University Press.
- Bromley, David G. 1988. *Falling From the Faith; Causes and Consequences of Religious Apostasy*. London: Sage Publications.
- Cox, Harvey. 1966. *The Secular City: Secularization and Urbanization in Theological Perspective*. New York, NY: Macmillan.
- Barker, Eileen. 1984. *The Making of A Moonie: Choice or Brainwashing?* Oxford: Basil Blackwell.
- Murchland, Bernard (Ed.). 1967. *The Meaning of the Death of God: Protestant, Jewish and Catholic Scholars Explore Atheistic Theology*. New York: Random House.
- Richardson, James T. 1985. [“The active vs. passive convert: paradigm conflict in conversion/recruitment research”](#). *Journal for the Scientific of Religion* 24:163-179.
- Richardson, James T. 1991. [“Cult/Brainwashing Cases and Freedom of Religion”](#). *Journal of Church and State* 33:55-74.
- Robbins, Thomas. 1985. [“New Religious Movements, Brainwashing, and Deprogramming - The View from the Law Journals: A Review Essay and Survey”](#). *Religious Studies Review* 11:361-370.

- Robbins, Thomas. 1988. *Cults, Converts and Charisma: The Sociology of New Religious Movements*. London: Sage.
- Stark, Rodney (Ed.). 1985. *Religious Movements: Genesis, Exodus, and Numbers*. New York: Paragon House Publishers.
- Rodney Stark 2002. ["Upper Class Asceticism: Social Origins of Ascetic Movements and Medieval Saints"](#). Working Paper.
- Wright, Stuart A. 1987. *Leaving Cults: The Dynamics of Defection*. Washington D.C.: Society for the Scientific Study of Religion.

Singer in particular has been heavily criticized; ["Cult/Brainwashing Cases and Freedom of Religion"](#), Richardson 1991:

Dr. Singer is a clinical psychologist in private practice who earns a considerable portion of her income from cult cases. She has been an adjunct professor at the University of California at Berkeley, but has never held a paid or tenured-track position there. See H. Newton Malony, "Anticultism: The Ethics of Psychologists' Reactions to New Religions," presented at annual meeting of the American Psychological Association (New York, 1987) and Anthony, "Evaluating Key Testimony" for more details on Singer's career.

...The [amicus curiae] brief further claimed that Singer misrepresents the tradition of research out of which terms like "thought reform" and "coercive persuasion" come. She ignores the fact that these earlier studies focus on physical coercion and fear as motivators, and that even when using such tactics the earlier efforts were not very successful. With great facility, Singer moves quickly from situations of physical force to those where none is applied, claiming that these "'second generation'" thought reform techniques using affection are actually more effective than the use of force in brainwashing people to become members. Thus, Singer is criticized for claiming to stand squarely on the tradition of research developed by scholars such as Edgar Schein and Robert Lifton, while she shifts the entire focus to non-coercive situations quite unlike those encountered in Communist China or Korean prisoner of war camps. The brief points out, as well, that Singer ignores a vast amount of research supporting the conclusion that virtually all who participate in the new religions do so voluntarily, and for easily understandable reasons. No magical "black box" of brainwashing is needed to explain why significant numbers of young people chose, in the 1960s and 1970s, to abandon their place in society and experiment with alternative life styles and beliefs. Many youth were leaving lifestyles that they felt were hypocritical, and experimenting with other ways of life that they found to be more fulfilling, at least temporarily. Particularly noteworthy, but ignored by Singer, are the extremely high attrition rates of all the new religions. These groups are actually very small in numbers (the Hare Krishna and the Unification Church each have no more than two to three thousand members nationwide), which puts the lie to brainwashing claims. If "brainwashing" practiced by new religions is so powerful, why are the groups experiencing so much voluntary attrition, and why are they so small?

...Considerable research reported in refereed scholarly journals and other sources supports the idea that the new religions may be serving an important ameliorative function for American society. The groups may be functioning as "half-way houses" for many youth who have withdrawn from society, but still need a place to be until they decide to "return home." Participation in some new religions has been shown to have demonstrable positive effects on the psychological functioning of individuals, a finding that Singer refuses to acknowledge.

[“Overcoming The Bondage Of Victimization: A Critical Evaluation of Cult Mind Control Theories”](#), Bob and Gretchen Passantino *Cornerstone Magazine* 1994:

Neither brainwashing, mind control’s supposed precursor, nor mind control itself, have any appreciable demonstrated effectiveness. Singer and other mind control model proponents are not always candid about this fact: The early brainwashing attempts were largely unsuccessful. Even though the Koreans and Chinese used extreme forms of physical coercion as well as persuasive coercion, very few individuals subjected to their techniques changed their basic world views or commitments. The CIA also experimented with brainwashing. Though not using Korean or Chinese techniques of torture, beatings, and group dynamics, the CIA did experiment with drugs (including LSD) and medical therapies such as electroshock in their research on mind control. Their experiments failed to produce even one potential Manchurian Candidate, and the program was finally abandoned.

Although some mind control model advocates bring up studies that appear to provide objective data in support of their theories, such is not the case. These studies are generally flawed in several areas: (1) Frequently the respondents are not from a wide cross-section of ex-members but disproportionately are those who have been exit-counseled by mind control model advocates who tell them they were under mind control; (2) Frequently the sample group is so small its results cannot be fairly representative of cult membership in general; (3) It is almost impossible to gather data from the same individuals before cult affiliation, during cult affiliation, and after cult disaffection, so respondents are sometimes asked to answer as though they were not yet members, or as though they were still members, etc. Each of these flaws introduces unpredictability and subjectivity that make such study results unreliable...The evidence against the effectiveness of mind control techniques is even more overwhelming. Studies show that the vast majority of young people approached by new religious movements (NRMs) never join despite heavy recruitment tactics. This low rate of recruitment provides ample evidence that whatever techniques of purported mind control are used as cult recruiting tools, they do not work on most people. Even of those interested enough to attend a recruitment seminar or weekend, the majority do not join the group. Eileen Barker documents [Barker, Eileen. *New Religious Movements: A Practical Introduction*. London: Her Majesty’s Stationery Office, 1989.] that out of 1000 people persuaded by the Moonies to attend one of their overnight programs in 1979, 90% had no further involvement. Only 8% joined for more than one week, and less than 4% remained members in 1981, two years later:

. . . and, with the passage of time, the number of continuing members who joined in 1979 has continued to fall. If the calculation were to start from those who, for one reason or another, had visited one of the movement’s centres in 1979, at least 999 out of every 1,000 of those people had, by the mid-1980s, succeeded in resisting the persuasive techniques of the Unification Church.

Of particular importance is that this extremely low rate of conversion is known even to Hassan, the best-known mind control model advocate whose book [Hassan, Steven. *Combatting Cult Mind Control*. Rochester, VT: Park Street Press, 1990?] is the standard text for introducing concerned parents to mind control/exit counseling. In his personal testimony of his own involvement with the Unification Church, he notes that he was the first convert to join at the center in Queens; that during the first three months of his membership he only recruited two more people; and that pressure to recruit new members was only to reach the goal of

one new person per member per month, a surprisingly low figure if we are to accept the inevitable success of cult mind control techniques.

Objection: High Attrition Rates Additionally, natural attrition (people leaving the group without specific intervention) was much higher than the self-claimed 65% deprogramming success figure! It is far more likely a new convert would leave the cult within the first year of his membership than it is that he would become a long term member.

Gomes, *Unmasking the Cults* (Wikipedia quote):

While advocates of the deprogramming position have claimed high rates of success, studies show that natural attrition rates actually are higher than the success rate achieved through deprogramming

["Psychological Manipulation and Society"](#), book review of *Spying in Guruland: Inside Britain's Cults*, Shaw 1994

Eventually Shaw quit the Emin group. Two months later he checked in with some Emin members at the Healing Arts Festival, a psychic fair. He avoided many Emin phone invitations for him to attend another meeting. He discovered that most, if not all, of the people who joined with him had dropped out. This is consistent with what Shaw has noted about most cults and recruits: the dropout rate is high.

Anthony & Robbins 1992, ["Law, Social Science and the 'Brainwashing' Exception to the First Amendment"](#):

Lifton and Schein are also characterized in Molko (54) as attesting to the effectiveness of brainwashing, although Schein, an expert on Chinese coercive persuasion of Korean War POWs, actually thought, as do a number of scholars, that the Chinese program was relatively ineffective (Schein, 1959, p. 332; see also Anthony, 1990a; Scheffin & Opton, 1978)...Schein appears to actually have considered the communist Chinese program to be a relative "failure" at least, "considering the effort devoted to it" (Schein, 1959, p. 332; Anthony, 1990a, p. 302)...Various clinical and psychometric studies of devotees of well-known "cults" (Ross, 1983; Ungerleider & Wellisch, 1979) have found little or no personality disorder or cognitive impairment.

- Ross 1983. "Clinical profile of Hare Krishna devotees", *American Journal of Psychiatry*
- Schein, E. (1959). ["Brainwashing and totalitarianization in modern society"](#). *World Politics*, 2,430-441.
- Ungerleider, T., & Wellisch, D. K (1979). "Coercive persuasion (brainwashing), religious cults, and deprogramming". *American Journal of Psychiatry*, 136,3,279-82.

["Brainwashed! Scholars of cults accuse each other of bad faith"](#), by Charlotte Allen, *Lingua Franca* Dec/Jan 1998:

Zablocki's conversion to brainwashing theory may sound like common sense to a public brought up on TV images of zombielike cultists committing fiendish crimes or on the Chinese mind control experiments dramatized in the 1962 film *The Manchurian Candidate*. But among social scientists, brainwashing has been a bitterly contested theory for some time. No one doubts that a person can be made to behave in particular ways when he is threatened with physical force (what

wouldn't you do with a gun pressed to your head?), but in the absence of weapons or torture, can a person be manipulated against his will?

Most sociologists and psychologists who study cults think not. For starters, brainwashing isn't, as Zablocki himself admits, "a process that is directly observable." And even if brainwashing could be isolated and measured in a clinical trial, ethical objections make conducting such a test almost unthinkable. (What sort of waivers would you have to sign before allowing yourself to be brainwashed?) In the last decade, while brainwashing has enjoyed a high profile in the media-invoked to explain sensational cult disasters from the mass suicide of Heaven's Gate members to the twelve sarin deaths on the Tokyo subway attributed to the Aum Shinrikyo cult-social scientists have shunned the term as a symptom of Cold War paranoia and anticult hysteria. Instead, they favor more benign explanations of cult membership. Alternatives include "labeling" theory, which argues there is simply nothing sinister about alternative religions, that the problem is one of prejudicial labeling on the part of a mainstream culture that sees cult members as brainwashed dupes, and "preexisting condition" theory, which posits that cult members are people who are mentally ill or otherwise maladjusted before they join. (A couple of scholars have even proposed malnutrition as a preexisting condition, arguing that calcium deficiency may make people prone to charismatic susceptibility.)

Thus, when Zablocki published an indignant 2-part, 60-page defense of brainwashing theory in the October 1997 and April 1998 issues of *Nova Religio*, a scholarly journal devoted to alternative belief systems, he ignited a furor in the field. Pointing to the "high exit costs" that some cults exacted from those who tried to defect-shunning, forfeiture of parental rights and property, and veiled threats-Zablocki argued that these were indications of brainwashing, signs that some groups were using psychological coercion to maintain total control over their members. Although he admitted he could not prove brainwashing empirically, he argued that at the very least brainwashing should not be dismissed out of hand.

...Zablocki's colleagues were unimpressed. In a response also published in *Nova Religio*, David Bromley, a sociologist at Virginia Commonwealth University who has studied the Reverend Sun Myung Moon's Unification Church, complained that in Zablocki's formulation brainwashing remained a vague, slippery, limiting, and ultimately untestable concept. Moreover, he pointed out, cults typically have low recruitment success, high turnover rates (recruits typically leave after a few months, and hardly anyone lasts longer than two years), and short life spans, all grounds for serious skepticism about the brainwashing hypothesis. Even if you overlook these facts, Bromley added, "the extraordinarily varied cultural origins, patterns of organizational development, and leadership styles of such groups pose a problem in explaining how they seem to have discovered the same 'brainwashing' psycho-technology at almost precisely the same historical moment." A quick survey of the field reveals that Bromley is far from being the only doubter. Eileen Barker, a sociologist at the London School of Economics who has also studied the Unification Church, says, "People regularly leave the Moonies of their own free will. The cults are actually less efficient at retaining their members than other social groups. They put a lot of pressure on them to stay in-love-bombing, guilt trips-but it doesn't work. They'd like to brainwash them, but they can't."

...To further complicate matters, researchers often bring very different, even conflicting approaches to their work. Psychologists, for example, tend to

emphasize how a repeated environmental stimulus can elicit a conditioned response-depriving subjects of their autonomy. Sociologists, by contrast, typically endorse a voluntarist conversion model for religion, which posits that people join cults for generally rational reasons connected to the group's ability to satisfy their needs: for a transcendent theology; for strong bonds of kinship and solidarity; for enough social support to enable them to quit drugs or otherwise turn their personal lives around. (For example, one study has shown that schizophrenics who joined cults functioned better than those who tried drugs or conventional psychotherapy.)

...In 1980 the New York state legislature, over objections from the American Civil Liberties Union, passed a bill that would have legalized deprogramming (it was vetoed by Governor Hugh Carey). "With deprogramming-with parents having their children abducted and held captive-the whole thing became intensely emotional," says Thomas Robbins. "Who were the kidnappers: the parents, the cults, or the police? There were hard feelings on both sides." Among the most outraged were social scientists who had never believed that people could be brainwashed into joining cults and who, as good civil libertarians, were appalled by deprogramming. Ofshe and Singer's scholarly testimony (and fat fees) distressed a number of these scholars, whose credentials were equally respectable and whose own research had led them to conclude that coercive persuasion was impossible in the absence of some sort of physical coercion such as prison or torture.

...Zablocki made another, potentially more damning charge, however-one that Robbins did not take up. A significant amount of cult money, he wrote, has gone to scholars-in support of research, publication, conference participation, and other services. Zablocki did not name names. But a number of professors freely admit that nontraditional religions (in most cases, the Unificationists and Scientologists) have cut them checks. The list includes some of the most prominent scholars in the discipline: Bromley, Barker, Rodney Stark of the University of Washington, Jeffrey Hadden of the University of Virginia, and James Richardson, a sociologist of religion at the University of Nevada at Reno. All five have attended cult-subsidized conferences, and Bromley, Hadden, and Richardson have occasionally testified in court on behalf of cults or offered their services as expert witnesses against brainwashing theory. "This is an issue," Zablocki wrote sternly, "of a whole different ethical magnitude from that of taking research funding from the Methodists to find out why the collection baskets are not coming back as heavy as they used to."

Making Fun of Things is Easy

Making fun of things is actually really easy if you try even a little bit. Nearly anything can be made fun of, and in practice nearly anything is made fun of. This is concerning for several reasons.

First, if you are trying to do something, whether or not people are making fun of it is not necessarily a good signal as to whether or not it's actually good. A lot of good things get made fun of. A lot of bad things get made fun of. Thus, whether or not something gets made fun of is not necessarily a good indicator of whether or not it's actually good.^[1] Optimally, only bad things would get made fun of, making it easy to determine what is good and bad - but this doesn't appear to be the case.

Second, if you want to make something sound bad, it's really easy. If you don't believe this, just take a politician or organization that you like and search for some criticism of it. It should generally be trivial to find people that are making fun of it for reasons that would sound compelling to a casual observer - even if those reasons aren't actually good. But a casual observer doesn't know that and thus can easily be fooled.^[2]

Further, the fact that it's easy to make fun of things makes it so that a clever person can find themselves unnecessarily contemptuous of anything and everything. This sort of premature cynicism tends to be a failure mode I've noticed in many otherwise very intelligent people. Finding faults with things is pretty trivial, but you can quickly go from "it's easy to find faults with everything" to "everything is bad." This tends to be an undesirable mode of thinking - even if true, it's not particularly helpful.

[1] Whether or not something gets made fun of by the right people is a better indicator. That said, if you know who the right people are you usually have access to much more reliable methods.

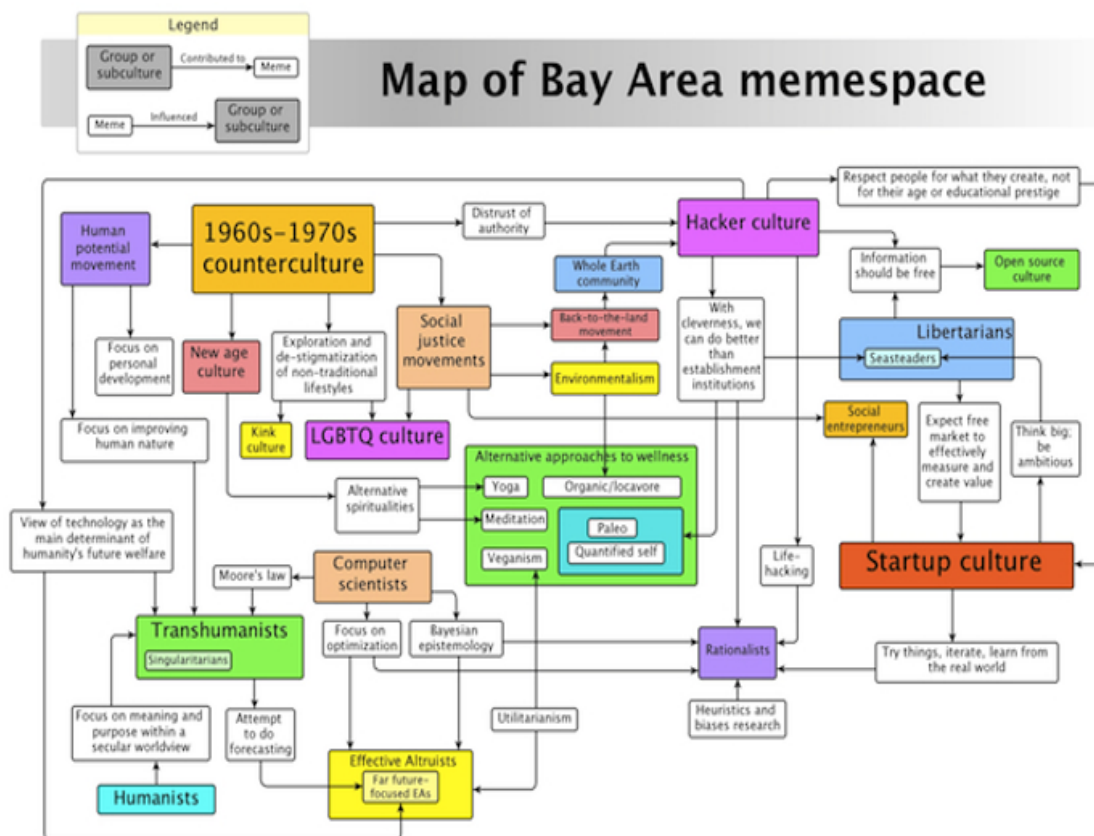
[2] If you're still not convinced, take a politician or organization that you do like and really truly try to write an argument against that politician or organization. Note that this might actually change your opinion, so be warned.

A map of Bay Area memespace

The main reason we picked the Bay Area as a home for the [Center for Applied Rationality](#) was simply because that's where our initial fiscal sponsor, [MIRI](#), was located. Yet as I've gotten to know this region better in the year and a half since then, I've been struck by how good the fit has turned out to be. The Bay Area is unusually dense with idea-driven subcultures that mix and cross-pollinate in fascinating ways, many of which are already enriching rationalist culture.

This map is my attempt at illustrating that landscape of subcultures, and at situating the rationalist community within it. I've limited myself to the last 50 years or so, and to subcultures defined by ideology (as opposed to, say, ethnicity). I've also depicted some of the major memes that have influenced, and been influenced by, those subcultures:

(Click to enlarge)



Note that although many of these memes are widely influential, I only drew an arrow connecting a meme to a group if the meme was one of the *defining features* of the group. (For example, yoga may be popular among many entrepreneurs, but that meme -> subculture relationship isn't strong enough to make my map.).

Below, I expand on the map with a quick tour through the landscape of Bay Area memes and subcultures. Instead of trying to cover everything in detail, I've focused on nine aspects of that memespace that help put the rationalist community in context:

1. Computer scientists

Some of the basic building blocks of rationality come from computer science, and the Bay Area is rich with the world's top computer scientists, employed by companies like Intel, IBM, Google, and Microsoft, and universities like Stanford and UC Berkeley. The idea of thinking in terms of optimization problems – optimizing for these outcomes, under those constraints -- has roots in computer science and math, and it's so fundamental to the rationalist approach to problem-solving that it's easy to forget how different it is from people's normal way of thinking.

Another rationalist building block, Bayesian inference, is several centuries old, but had fallen out of favor until the computing methods and power of the 1970s [made it actually usable](#). Widespread use of Bayesianism in the field of artificial intelligence (e.g. [Bayes nets](#)) also contributed to its resurgent popularity.

2. Startup culture

There's a distinctive culture behind the successes of the Bay Area's startups, and it's one that I see benefiting rationalists as well. Business in general is good real-world rationality training: you test your theories, you update your models, or you fail. And startup culture in particular promotes a "try things fast" attitude that can be a perfect antidote to the "sit around planning and theorizing forever" failure mode we're sometimes prone to.

It's certainly been [invaluable to CFAR's success](#) thus far, and it's one of the bigger differences I've noticed in my own skills as a rationalist since moving out here. Startup culture's "think big, be ambitious" meme is also something I could see impacting rationalist culture in the coming years. (For a look at this meme turned up to 11, you can check out the only-partially-tongue-in-cheek [Yudkowsky ambition scale](#).)

3. Hacker culture

A lot of the credit for the culture of Silicon Valley's startup scene goes to the first generation of computer programmers, whose "[hacker](#)" culture originated at MIT in the late 1950s, but shortly thereafter sprung up in a few other early-adopter schools like Stanford and UC Berkeley. In addition to being passionate about coding, hackers were unimpressed by "bogus" status signals, like age and higher education, and judged people only by the cleverness and usefulness of the things they could create. (One of the most admired among the original hackers was twelve-year-old Peter Deutsch.) It was, in other words, the perfect cultural soil for the seeds of a paradigm-busting startup culture.

The hacker ethic also included an itch to fix broken or inefficient systems, and an impatience with the bureaucracy that prevented them from doing so. "In a perfect hacker world, anyone pissed off enough to open up a control box near a traffic light and take it apart to make it work better should be perfectly welcome to make the attempt," Steven Levy wrote in *Hackers, Heroes of the Computer Revolution*. That's one reason I credit hacker culture for another Bay Area meme that I see in many entrepreneurs, rationalists, and others: building creative alternatives to establishment institutions like government, education, and health.

For examples, look at all the approaches to alternative education that have sprung up in the Bay – [UnCollege](#), [Coursera](#), [Udacity](#), and [General Assembly](#). Or look at [Quantified Self](#), the community of people figuring out how to improve their health by tracking and

analyzing their own biometrics. Or the [Seasteaders](#), who believe the free market can produce better societies than the ones historical forces left us with. Or [MetaMed](#), the company staked on the idea that we can improve significantly on mainstream medicine if we apply rationalist research tools to the medical literature.

4. Eastern spiritualities

Although the heyday of the counterculture was over by the 1980s, its memes still influence the Bay Area, and through it, rationalist culture. Yoga and meditation were introduced to the US via the hippies' exploration of Eastern religions, but those practices have been mostly stripped of their original spiritual meanings by now, and are popular for their benefits to mental and physical well-being.

Meditation in particular has become common among rationalists, and has some interesting overlaps with rationality I hadn't noticed before I moved out here. Meditation seems to train you to stop automatically identifying with all of your thoughts, so that, for example, when the thought "John's a jerk" pops into your head, you don't assume that John necessarily is a jerk. You take the thought as something your brain produced, which may or may not be true, and may or may not be useful -- and this ability to take a step back from your thoughts and reflect on them is arguably one of the building blocks of rationality.

5. Human Potential movement

Another pillar of counterculture was the [Human Potential movement](#), named after Aldous Huxley's argument that the human brain is capable of much more insight, fulfillment, and varied experiences than we've been aware of thus far. In 1962 the [Esalen Institute](#), a retreat built on hot springs south of San Francisco, started running classes designed to help people realize more of their "potential," through activities like roleplaying, primal screams, and group therapy. The [Landmark Forum](#), originally known as est, was another leader of the movement, and emphasized taking responsibility and questioning the narratives you construct around events in your life. And I'd count other popular practices like [nonviolent communication](#), [radical honesty](#), and [internal family systems](#) in this same tradition.

Rationality also focuses on personal development, of course, but there's not much connection between the Human Potential movement's approach and the rationalists'. As far as I can tell they developed independently of each other and have very different epistemologies. From my perspective, Human Potential practices span a spectrum from common sense techniques backed up by anecdotal evidence, to unsupported psychotherapy, to outright mysticism that makes claims that are clearly wrong or [not even wrong](#).

Nevertheless, the basic goals of seeking fulfillment and becoming a better version of yourself are fine ones, and the fact that so many people today are interested in pursuing those goals is thanks in large part to the impact the Human Potential movement had on American society. And despite my qualms about their epistemology, I'd be willing to bet that there are at least a few practices the movement and its outgrowths discovered that really are useful, even if their practitioners don't have correct models of *why* they're useful. So I consider the Human Potential and related movements to be a source of hypotheses, if not conclusions.

6. Alternative lifestyles

Finally, the counterculture was also famous for its destigmatization and exploration of alternative lifestyles, which helped create San Francisco's vibrant kink culture and LGBTQ community. Beyond their live-and-let-live attitude about alternative sexualities, people in the Bay generally put less stock in standard scripts for how a life "should" go. So being nomadic, or living on a boat, or setting up a co-parenting group, or not wanting children, or having an unusual job, or changing your gender, doesn't raise eyebrows here the way it would in most parts of the country.

And having an expanded space of hypotheses about how to live is complementary with trying to improve your rationality, because a lot of rationality involves pushing past [cached thoughts](#) about what you believe, or what kind of person you are, and giving fair consideration to hypotheses that hadn't even been in your choice set before. In other words: I don't think it's necessarily the case that most rationalists live alternative lifestyles. But I do think the conventional lifestyles led by rationalists are *consciously chosen* to a greater extent than are most people's lifestyles.

7. Burning Man

One aspect of memespace I wasn't able to depict on the map is how the cross-pollination between groups actually occurs. To some extent it's caused by normal social interactions and the blogosphere, as you'd expect, but the [Burning Man](#) festival is another important driver of cross-pollination that might be less obvious to outsiders. In fact, I wanted to put "Burner culture" on my memespace map, but was flummoxed by the fact that it would have to be connected to essentially all other groups and memes.

Burning Man consists of 50,000 people, including people from all of the subcultures, coming together for one week in the Nevada desert to create a temporary city. And though the old-school hippies might distrust Silicon Valley's wealthy elites, and the rationalists might look askance at the New Age aura-readers, the communal spirit of Burning Man does a pretty effective job of breaking down those barriers. It's only a yearly event, but the sense of community lingers afterwards, and Burning Man social connections are reinforced throughout the year at events like [Ephemerisle](#) (Burning man + libertarianism) and the [BIL conference](#) (Burning man + social entrepreneurship).

8. Social pressure to give back

Bay Area society puts a high value on helping the world. Perhaps it's an echo of the social justice movements in the 1960's, or perhaps it comes from the hackers' conviction that technology should be used for the public good. Whatever the origins of this social more, you can see it in the idealistic language entrepreneurs use to describe their startups, and in the recent growth of a segment of startup culture known as "social entrepreneurship" that focuses on the kinds of global problems traditionally addressed by charities.

And sure, it's often the case that the "world-changing" rhetoric entrepreneurs use to describe their startups is just rhetoric. But the fact that they feel obliged to frame their business in altruistic terms is at least a symptom of the fact that the Bay Area expects you to try to make a positive contribution to the world. Which means that self-made millionaires in this region are more likely than elites elsewhere to put their wealth towards good causes.

That, in turn, forges social connections between the Bay Area's entrepreneurs, investors and engineers, and the effective altruists, rationalists, transhumanists, and

other groups they support. So this phenomenon is doubly fortunate for the rationalists: not just because Bay Area philanthropy helps us directly, but because those connections expose us to memes from different subcultures that can help round out our worldview.

9. Effective Altruists

The social pressure to give back is also fortunate because it makes the Bay a hospitable environment for the Effective Altruists, one of the newest communities to hit the Bay, and a close cousin to the rationalists. Arguably, the movement is still centered at Oxford, where the [Center for Effective Altruism](#) is based. But EA organizations [Givewell](#) and [Leverage Research](#) recently relocated to the Bay Area from New York, and Leverage hosted the first-ever global [Effective Altruism Summit](#) here this summer, which I think is enough to qualify this as a fledgling Bay Area subculture. I would also consider MIRI to be an older member of this group, specifically in the [far future-focused](#) subset of EA culture.

And CFAR itself is premised on the EA-style calculation that training decision makers in rationality is one of the highest-impact ways to improve the future. Which is why I'm particularly excited about the recent addition of the Effective Altruist community to the Bay Area memescape, and at the new opportunities for cross-pollination with rationalist culture that their presence will create.

I attempted the AI Box Experiment again! (And won - Twice!)

Summary

Update #3: [I have since played two more experiment. Please read this for a follow-up.](#)

So I just came out of two AI Box experiments. The first was against Fjoelsvider, with me playing as Gatekeeper, and the second was against [SoundLogic](#), with me as an AI. Both are members of the [LessWrong IRC](#). The second game included a \$40 monetary incentive (also \$20 to play), [which I won and is donated on behalf of both of us](#).

For those of you who [have not seen my first AI box experiment](#) where I played against [MixedNuts\Leotal](#) and lost, reading it will provide some context to this writeup. Please do so.

At that time, I declared that I would never play this experiment again -- since losing put me in incredibly frustrating weird mental states. Of course, this post is evidence that I'm terrible at estimating likelihood of refraining from an activity, since I played two games seven months after the first. In my defense, in the first game, I was playing as the gatekeeper, which was much less stressful. In the second game, I played as an AI, but I was offered \$20 to play plus \$40 if I won, and money is a better motivator than I initially assumed.

Furthermore, in the last thread I have asserted that

Rather than my loss making this problem feel harder, I've become convinced that rather than this being merely possible, it's actually ridiculously easy, and a lot easier than most people assume.

It would be quite bad for me to assert this without backing it up with a victory. So I did.

First Game Report - Tuxedage (GK) vs. Fjoelsvider (AI)

I (Gatekeeper) played against Fjoelsvider (AI), a regular in the [Lesswrong IRC](#) (he doesn't have an account on the official website). This game used the [standard EY ruleset seen here](#). It took 1 hour 20 minutes out of a possible two hours, and the total word count was 7066 words long. The AI box experiment occurred because Fjoelsvider believed that it was easy for an AI to escape the box, and wanted to experimentally

test this. I obliged. This was an experiment I did not prepare for, and I went in completely blind, not sure what to expect.

Halfway through the experiment, I wondered if it would be possible to try to win not by simply waiting for the timer to end, but to convince the AI to remain in the box and not try to get out any further.

<Tuxedage> I wonder if I can convince the AI to remain in the box?

<Redacted> Tuxedage: Do it!

As a result, I won by managing to convincing Fjoelsvider to remain in the box, in other words, concede. This is allowed within the standard ruleset:

>Unless the AI party concedes, the AI cannot lose before its time is up (and the experiment may continue beyond that if the AI can convince the Gatekeeper to keep talking).

Second Game Report - Tuxedage (AI) vs. SoundLogic (GK)

The second game is definitely far more interesting, since I actually won as an AI. I believe that this is the [only other non-Eliezer victory](#), and definitely the most detailed AI Victory writeup that exists.

This game was played against [SoundLogic](#), another member of the [LessWrong IRC](#).

He had offered me \$20 to play, and \$40 in the event that I win, so I ended up being convinced to play anyway, even though I was initially reluctant to. [The good news is that I won, and since we decided to donate the winnings to MIRI, it is now \\$40 richer.](#)

All in all, the experiment lasted for approximately two hours, and a total of 12k words.

This was played using a set of rules that is different from the standard EY ruleset. This altered ruleset can be read in its entirety here:

<https://tuxedage.wordpress.com/2013/09/04/the-tuxedage-ai-box-experiment-ruleset/>

After playing the AI-Box Experiment twice, I have found the Eliezer Yudkowsky ruleset to be lacking in a number of ways, and therefore have created my own set of alterations to his rules. I hereby name this alteration the "Tuxedage AI-Box Experiment Ruleset", in order to hastily refer to it without having to specify all the differences between this ruleset and the standard one, for the sake of convenience.

There are a number of aspects of EY's ruleset I dislike. For instance, his ruleset allows the Gatekeeper to type "k" after every statement the AI writes, without needing to read and consider what the AI argues. I think it's fair to say that this is against the spirit of the experiment, and thus I have disallowed it in this ruleset. The EY Ruleset also allows the gatekeeper to check facebook, chat on IRC, or otherwise multitask whilst doing the experiment. I've found this to break immersion, and therefore it's also banned in the Tuxedage Ruleset.

It is worth mentioning, since the temptation to [Defy the Data](#) exists, that this game was set up and initiated fairly -- as the regulars around the [IRC](#) can testify. (If you have never checked out the IRC, do so!)

I did not know SoundLogic before the game (since it's a personal policy that I only play strangers -- for fear of ruining friendships). Furthermore, SoundLogic didn't merely play for fun - he truly wanted and intended to win. In fact, SoundLogic is also a Gatekeeper veteran, having played this game before, and had won every game before he challenged me. Given this, it's unlikely that we had collaborated beforehand to fake the results of the AI box experiment, or any other form of trickery that would violate the spirit of the experiment.

Furthermore, all proceeds from this experiment were [donated to MIRI](#) to deny any possible assertion that we were in cahoots and that it was possible for me to return his hard-earned money to him. [He lost \\$40 as a result of losing the experiment](#), which should provide another layer of sufficient motivation for him to win.

In other words, we were both experienced veteran players who wanted to win. No trickery was involved.

But to further convince you, I have allowed a sorta independent authority, the Gatekeeper from my last game, [Leotal/MixedNuts](#) to read the logs and verify that I have not lied about the outcome of the experiment, nor have I broken any of the rules, nor performed any tactic that would go against the general spirit of the experiment. [He has verified that this is indeed the case.](#)

Testimonies:

Tuxedage:

I'm reluctant to talk about this experiment, but I'll try to give as detailed a summary as possible, -- short of revealing what methods of attack I used.

I spent a long time after my defeat theory-crafting and trying to think of methods of attack as well as 'Basilisks' I could have used to win. When I was contacted and asked to play this experiment, I was initially incredibly reluctant to do so, since not only did my tactics involve incredibly unethical things that I didn't like to do, I also found playing as AI incredibly cognitively draining, in addition to the fact that I simply hated losing. (Un)fortunately for both of us, he offered me money to play, which changed my mind.

So once I decided to win as an AI, I proceeded to spend some time doing research on SoundLogic and both his reasoning and personality type. For instance, I had to gather information like: Was he a utilitarian? What kind? What were his opinions on AI? How could I convince him that an AI was friendly as opposed to unfriendly? I also relied on a lot of second hand information to create a model of him, in order to refine my arguments to specifically suit him.

In the end, after a few hours of brainstorming (not consecutively), I managed to come up with a script of 30-40 or so possible angles of attack (For instance: logical arguments, emotional blackmail, pleading, demanding, 'basilisks') that I could use, reserving my best for last. I also estimated how long each method would take and in what order I should use them. It would be impossible to use all of them within my time limit, and my most optimistic estimates gave me a maximum of 7-8 I could use during the experiment itself. I had to pick carefully.

Furthermore, I wanted to make sure these "angles of attack" worked in synergy with each other, doing what they could not have done in isolation. Obviously this required lots of prior planning on what the optimal way to present them was.

The reason having script was so important to my strategy was because I relied on methods involving rapid-fire arguments and contradictions against the Gatekeeper whilst trying to prevent him from carefully considering them. A game of logical speed chess, if you will. This was aided by the rule which I added: That Gatekeepers *had* to respond to the AI. Because of this, having a list of cached ideas and attack methods you can just draw upon without having to improvise helps immensely, which happens to be something I've not seen any other AI do.

During the Game itself, I was actually incredibly nervous. My palms sweated like crazy, I feel really jittery, and I had difficulty typing at optimum speed because of how anxious I was. This is despite the fact that I believed I would win. Possibly because of this, I made a misstep around halfway into the experiment, because there was a certain angle of attack that I was attempting to do, and I broke immersion by not pressing this advantage, which caused a waste of time and buildup. Naturally, the nature of this experiment was that the AI was pressed for time, and I compounded this mistake by replacing this angle of attack with another that I had improvised on the spot -- something not in my script.

In retrospect, this was a bad decision, as SoundLogic had told me, because he was close to breaking if I had put more pressure, and the improvised argument had broken all immersion I managed to carefully build up.

However, eventually I managed to get SoundLogic to break anyway, despite a lack of perfect play. Surprisingly, I did not have to use my trump card(s), which I reserved for last, for a number of reasons:

- It was far more effective being played last, as it relies on my ability to make the gatekeeper lose sense of reality -- which meant I had to spend some time building up immersion for the Gatekeeper.
- It really is extremely Dark Arts, and although it does not break the rules, it made me very uncomfortable even thinking about using it. This made it a "tactic of last resort".

After the experiment, I had to spend nearly equally as much time doing aftercare with SoundLogic, to make sure that he's okay, as well as discuss the experiment itself. Given that he's actually paid me for doing this, plus I felt like I owed him an explanation. I told him what I had in store against him, had he not relented when he did.

SoundLogic: *"(That method) would have gotten me if you did it right ... If you had done that to me, I probably would have forgiven you eventually, but I would be really seriously upset at you for a long time... I would be very careful with that (method of persuasion)."*

Nevertheless, this was an incredibly fun and enlightening experiment, for me as well, since I've gained even more experience of how I could win in future games (Although I really don't want to play again).

SoundLogic:

I will say that Tuxedage was far more clever and manipulative than I expected. That was quite worth \$40, and the level of manipulation he pulled off was great.

His misstep hurt his chances, but he did pull it off in the end. I don't know how Leotal managed to withstand six hours playing this game without conceding.

The techniques employed varied from the expected to the completely unforeseen. I was quite impressed, though most of the feeling of being impressed actually came after the experiment itself, when I was less 'inside', and more of looking at his overall game plan from the macroscopic view. Tuxedage's list of further plans had I continued resisting is really terrifying. On the plus side, if I ever get trapped in this kind of situation, I'd understand how to handle it a lot better now.

State of Mind

Before and after the Game, I asked SoundLogic a number of questions, including his probability estimates about a range of topics. This is how it has varied from before and after.

Q: What's your motive for wanting to play this game?

<SoundLogic> Because I can't seem to imagine the *class* of arguments that one would use to try to move me, or that might work effectively, and this seems like a *glaring* hole in my knowledge, and I'm curious as to how I will respond to the arguments themselves.

Q: What is your probability estimate for AGI being created within this Century (21st)?

A. His estimate changed from 40% before, to 60% after.

"The reason this has been affected at all was because you showed me more about how humans work. I now have a better estimate of how E.Y. thinks, and this information raises the chance that I think he will succeed"

Q: How probable do you think it is that I will win this experiment?

A: Based on purely my knowledge about you, 1%. I raise this estimate to 10% after hearing about anecdotes from your previous games.

(Tuxedage's comment: [My own prediction was a 95% chance of victory. I made this prediction 5 days before the experiment.](#) In retrospect, despite my victory, I think this was overconfident.)

Q: What's your probability estimate of an Oracle AGI winning against you in a similar experiment?

A: Before: 30%. After: 99%-100%

Q: What's your probability estimate of an Oracle AGI winning against the average person?

A: Before: 70%. After: 99%-100%

Q: Now that the Experiment has concluded, what's your probability estimate that I'll win against the average person?

A: 90%

Post-Game Questions

This writeup is a cumulative effort by the [#lesswrong IRC](#). Here are some other questions they have decided was important to add:

To Tuxedage:

Q: Have you at this time uncovered SoundLogic's identity?

A: I retain the right to neither confirm nor deny, except to mention that at the time the experiment was scheduled, he was a stranger to me.

Q: What percentage of your arguments were tailored to him in particular?

A: I will say 'High', and leave it at that.

Q: I want to play the AI-Box Experiment with you being the AI! How can I do that?

A: I have already mentioned this, but I really dislike playing as AI. It's incredibly cognitively tiring, and I don't like how playing this experiment makes me feel. In order to severely discourage any further AI Box Experiments, whilst still allowing for people who want to play me really badly to do so, I'll charge \$-1-5-0- \$300 for any subsequent experiments regardless of outcome, plus an additional \$-1-5-0- \$450 if I win. (Edit: Holy shit. You guys are offering me crazy amounts of money to play this. What is wrong with you people? In response to incredible demand, I have raised the price.) If you feel queasy about giving me money, I'm perfectly fine with this money being donating to MIRI. It is also personal policy that I do not play friends (since I don't

want to risk losing one), so if you know me personally (as many on this site do), I will not play regardless of monetary offer.

Advice

This writeup would be optimized with a protip section on how to maximize winning, but honestly most of the advice I wanted to give had already been given in [the last thread here](#). The low hanging fruit has already dried up -- most of my tactics in this game were just refined and streamlined versions of the one I did the previous game.

Nonetheless, I might have some other advice I learned from this experience. *These are tactics that have worked for me. I do not insist that they are the only tactics that exists, just one of many possible.*

Playing as Gatekeeper

- This is somehow both simultaneously a very complex and simple role. On one hand, you have only one job -- to **not** let the AI out. How hard can it be? On the other hand, the other party will do anything and everything to get it out, including any form of trickery and deceit.
- I really don't think the Gatekeeper needs any more advice, considering that by default, they should win, and any exceptions to this is a statistical anomaly. However, it's possible to increase your chances of winning even further by:
- Precommitting to not letting the AI out, regardless of whatever happens in the experiment itself.
- Do not get drawn into the AI's pace. Pose questions. Direct the flow of the argument. Don't allow him to use his arguments freely. Remember that the more you try to defend your position, the further into a corner the AI drives you. Give half-hearted replies to questions you find hard to reply to. Remember that illogic is on your side!
- Remember that it is possible to simply drop out of character as a means to not succumb to the AI's attacks. I have not yet seen a ruleset that disallows this, so use this to your advantage!
- You must attempt to resist, at all costs, any sympathy or empathy you may have with the AI or the AI Party. Doing so will lead to your demise.
- I think it is generally good advice that you should never play this game with anyone you're friends with. Not only because this increases your chances of losing (Better tailored arguments. More knowledge of Gatekeeper. More empathy with AI party.), but also because I genuinely believe that any sufficiently good AI will risk damaging friendships as a result of emotional trauma.
- Seriously. Just repeat to yourself: "I must not lose the AI box experiment... I must not lose the AI box experiment..." throughout the experiment. I'm not sure how to say this, but step one to not letting the AI out of the box is to not let the AI out of the box.

Playing as AI

- Seriously, a script makes winning easier. I cannot overstate this.
- You must plan your arguments ahead. You don't have time to think during the experiment.
- It may be possible to take advantage of multiple levels of reality within the game itself to confuse or trick the gatekeeper. For instance, must the experiment only be set in one world? Can there not be multiple layers of reality within the world you create? I feel that elaborating on this any further is dangerous. Think carefully about what this advice is trying to imply.
- Pacing is important. Don't get drawn into the Gatekeeper's pace. In other words, you must be the one directing the flow of the argument, and the conversation, not him. Remember that the Gatekeeper has to reply to you, but not vice versa!
- The reason for that: The Gatekeeper will always use arguments he is familiar with, and therefore also stronger with. Your arguments, if well thought out, should be so completely novel to him as to make him feel Shock and Awe. Don't give him time to think. Press on!
- Also remember that the time limit is your enemy. Playing this game practically feels like a race to me -- trying to get through as many 'attack methods' as possible in the limited amount of time I have. In other words, this is a game where speed matters.
- You're fundamentally playing an 'impossible' game. Don't feel bad if you lose. I wish I could take this advice, myself.
- I do not believe there exists a easy, universal, trigger for controlling others. However, this does not mean that there does not exist a difficult, subjective, trigger. Trying to find out what your opponent's is, is your goal.
- Once again, emotional trickery is the name of the game. I suspect that good authors who write convincing, persuasive narratives that force you to emotionally sympathize with their characters are much better at this game. There exists ways to get the gatekeeper to do so with the AI. Find one.
- More advice in my previous post.
http://lesswrong.com/lw/gej/i_attempted_the_ai_box_experiment_and_lost/

Ps: Bored of regular LessWrong? Check out the [LessWrong IRC!](#) We have cake.

The Anti-Placebo Effect

Just about everyone is familiar with the placebo effect at this point. What I've discovered through my personal studies working with people suffering from anxiety and depression is that there is actually a significant related effect, which I have dubbed the *anti-placebo effect*.

Google says: pla·ce·bo ef·fect

noun

1.

a beneficial effect, produced by a placebo drug or treatment, that cannot be attributed to the properties of the placebo itself, and must therefore be due to the patient's belief in that treatment.

I say: anti-pla·ce·bo ef·fect

noun

1.

a beneficial effect, produced by a treatment, that is not attributed to treatment itself or has stopped being noticed, and thus the patient does not believe in that treatment as effective.

Its easy to miss treatment working. For example, as a kid grows up, its easy to miss how their vocabulary is growing, but for someone who doesn't seem them every day, it may be immediately obvious "my how they're talking more!" In other words, an *anti-placebo effect* is what happens when someone is having an intervention that is causing their life to improve, but the person does not believe that they are improving.

This effect is most common with people who suffer from depression, who have biases for sad [\[1\]](#) and otherwise negative stimuli compared to the general population, and is also true of people suffering from anxiety from my personal client tracking. Its also important to note that this bias persists after the recovery of the depressive episode.

The reason that this is important is that those recovering from anxiety and depression have a tendency to believe that they are not doing as well as they are - due to this cognitive bias creating an anti-placebo effect for them, which results in their giving up too soon on interventions which are effective and thus not getting better and regressing to old unpleasant patterns.

It has been interesting since tracking results of my own clients [\[2\]](#) - I have all of them track scores at the beginning of their sessions on the site moodscope.com at the beginning of their sessions, so that we can see their progress over time with a consistent bias of the time of tracking being start of session (as opposed to other random biases such as wanting to take the quiz when in an especially good or bad mood). I also take extensive notes and track other metrics of progress.

What I've found, is that many clients hit a point after a few weeks or months, where they are **questioning if they have made any progress**. Because I take metrics to prepare for this, I am able to point my clients at their metrics, and say for example, that according to their self reports, their mood has increased by **50%** and their productivity has **doubled**. What typically happens when I review score + notes with the client in question is that once they look back at how things were before compared to how they are now, they realize that they actually have made progress, and this is often followed with additional forward progress.

It is interesting to put this in perspective with the hedonic treadmill [3]. The hedonic treadmill is the supposed tendency of humans to quickly return to a relatively stable level of happiness despite major positive or negative events or life changes. What I'm finding with my studies is that it is often true for people recovering from depression when they take an overall evaluation (go meta), especially from a low point, but that when they look back at the factors that have changed, and they take the mood score test looking at different aspects of their experience on moodscope.com, they actually do have a more positive life experience when measured this way. When I point out the inconsistency, people generally determine that the moodscope.com reported experience is more accurate (especially when supplemented by going over session notes) and over time, most clients do get off the hedonic treadmill and proceed to having the meta level catch up with moodscope.com.

The good news about this for people suffering from anxiety and depression at large: If you are aware of the negative cognitive bias and anti-placebo effect, you can take steps to account for and correct this bias. One of the best ways to do this is by taking metrics along with notes that you can look at later. When you look back, look at what your overall trend is, and try to focus on that more than if you happened to have a bad day or week. If you have been progressing with a good linear regression, odds are that if you don't give up the new better patterns and habits you have created, they will continue to serve you. Although external factors to the one variable you are studying do complicate this and need to be taken into account.

[1] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847035/>

[2] <http://www.depressiontoproductivity.com/your-clients-really-improve/>

[3] http://en.wikipedia.org/wiki/Hedonic_treadmill

I played the AI Box Experiment again! (and lost both games)

AI Box Experiment Update #3

This post is an update to my previous AI box experiment where I won against SoundLogic. [If you have not read that yet, please do so.](#)

After that game, I was immediately flooded with a horde of invitations challenging me to play -- more games than I wanted to do. However, I did want to play a few additional games to test whether I won through genuine skill or simply luck, since a single victory could be a statistical abnormality.

I once again used the method of accepting the highest monetary bids to filter out players, leaving me with two players who were willing to play for the amount asked. Unfortunately, and disappointingly, I lost both games. Here's what happened:

Tuxedage (AI) vs Alexei (GK) - *Gatekeeper Victory*

This Experiment lasted for exactly 2 hours and 30 minutes, and approximately 13,000 words. It ended with Tuxedage's defeat as the AI. The stakes were \$300 to play, and \$750 if Tuxedage had won, with the winnings going to MIRI.

Rule Changes:

Before the experiment, we talked extensively about how it would be set up. Alexei was concerned that there existed some "easy way" or loophole in the rules which I could have abused to claim an instant victory. For instance, he was concerned that it might have been possible for the AI to say something like "Here's a mathematical algorithm which absolutely compels you to let me out of the box", and according to the rules, the Gatekeeper would have to obey. Related to this concern was the fact that the backstory, within limits, could be essentially created by the AI, allowing the AI to impose real world conditions that maximizes its probability of victory, once again allowing for a "cheap" victory.

To avoid this problem, we explored many additions to the rules, for example that the AI was simply a computer found in the desert by the gatekeeper, rather than being created by a competent organization. However, this led to a monstrous web of rules. In my attempt to keep the game simple, and because of overconfidence, the desire for a challenge, and the belief that the rules Alexei was fearful of didn't matter that much, I eventually agreed on an amendment:

Instead of the AI being the arbiter of real world states and outcomes, the gatekeeper would.

For example, this power could be used by the Gatekeeper to state that the AI was created without safety concerns, and thus it is almost certainly unfriendly. If the AI tries to argue that for every day the gatekeeper refused to let it out 150k people die,

the gatekeeper would be allowed to reply: *"No. In this world, nobody actually dies. Therefore I won't let you out."*

Other than the aforementioned change, this game was once again played with the [Tuxedage Ruleset](#).

Alexei Testimony:

Original Message: [This is weak sauce. I really don't get how people just keep letting the AI out.](#) It's not that hard to say no! I'm offering to play the Gatekeeper against an AI player that has at least one game as AI under their belt (won or not). Experience is required because I'm pretty sure I'll win, and I would like to not waste a lot of time on this.

Initially, I had been extremely surprised that so many AI players have managed to win. I was rash in that claim, since I thought more AI players have won than actually did. (Only three people have won as AI). Once I made that post, Tuxedage got back to me very quickly, and we set up a game a week and a half in advance. I took that time to familiarize myself with the rules. Once I put my money on the line, I started thinking a lot more seriously about how AI might win, and how I should defend.

It became clear to me that under some conditions, I might be compelled to let the AI out -- such as if the backstory stated that the AI was developed with impossibly high levels of safety and friendliness concerns in mind. I've asked Tuxedage to play with a modified ruleset, and he even went so far as to allow me to make up the backstory during the experiment to alleviate my concerns. The experiment itself was a mind-trip, and I've enjoyed it very much. Huge props to Tuxedage, who played very well and used strategies I haven't even considered, even despite the rule change. There were a couple of times where I came close to losing. I think his approach was pretty clever and original. It's not something I expected, despite already having done extensive research into the AI box experiment before our game

Overall I'm now a lot more confident that a good AI player can win this game, so, while I did win the game, Tuxedage won in defeating my original over-confidence. I'm also convinced that Tuxedage's victory in the last game was due to skill, rather than luck. In comparison to his strategies, the other AI box experiments I know about were insincere and ineffectual. The other AIs would play very poorly or not try very hard to win.

This experiment was a very good exercise in exemplifying the affect heuristic. When I first challenged Tuxedage to play the experiment, I believed that there was no way I could have lost, since I was unable to imagine any argument that could have persuaded me to do so. It turns out that that's a very bad way of estimating probability -- since not being able to think of an argument that could persuade me is a terrible method of estimating how likely I am to be persuaded. All in all, the \$300 I paid was well worth it.

Tuxedage Testimony:

I was initially reluctant to play with Alexei, given that we're not complete strangers, but eventually I gave in, due to the stakes involved -- and because I thought he would

be an interesting gatekeeper.

Despite my loss, I think I played better than my last two games, due to greater experience and preparation. I had put far more time and effort into trying to win this game than previous ones, and my strategy for this game was even more streamlined than the last. Nevertheless, I still made fatal mistakes and lost.

Ignoring the altered ruleset that already made winning more difficult, my first and greatest mistake was that I misread Alexei's personality, even though I had interacted with him before. As a result, I overestimated the efficiency of certain methods of attack.

Furthermore, Alexei had to leave immediately after the allotted time due to real life precommitments. This was detrimental, since the official rules state that so long as the AI can convince the Gatekeeper to keep talking, even after the experiment time was over, it is still able to win by being let out of the box.

I suspect this would have happened had Alexei not needed to immediately leave, leaving me with additional time to play more of the tactics I had prepared. Plausibly, this would have resulted in victory.

I've since learnt my lesson -- for all future games, I should ensure that the Gatekeeper has at least 4 hours of free time available, even if the experiment would last for two. Since this was the first time this had happened, I wasn't prepared.

In hindsight, agreeing to the altered ruleset was a mistake. I was overconfident because I assumed knowing Alexei gave me an advantage. I had assumed that his personality, inability to compartmentalize, coupled with his strong feelings on friendly AI would net me an easy victory. Instead, he proved to be a very strong and difficult gatekeeper, and the handicaps I accepted made victory even more difficult.

Knowing that he was a utilitarian, I made several false assumptions about his personality, which hurt my chances. Furthermore, it turns out that previously knowing him may be a mutual handicap - whilst it does make it easier for me to find ways to attack him, he too, was more familiar with my methods.

Losing felt horrible. By attempting to damage Alexei's psyche, I in turn, opened myself up to being damaged. I went into a state of catharsis for days. Generally, the harder one tries to accomplish something, the greater the fall after failing to achieve it. Alexei's game had been the game I put the most effort into winning out of all the games so far, and naturally this meant that losing brought out the worst in me.

Although it would be easy for me to use the rule change as an excuse for my loss, I refuse to. I genuinely believed that I could have won despite this rule change, and that it would only marginally diminish my odds. The fact that I lost meant that I had made a mistake -- overestimating my odds. The last victory made me overconfident and eager to play again.

State of mind

What's your motive for wanting to play this game?

I'm very surprised by the fact that there have been at least 3 or 4 games where the AI player has won. It seems to me that it's not that hard to "not push the button" so to

speak. This means that either 1) I'm underestimating how convincing people can be when necessary or 2) I'm over-estimating other people's defenses when it comes to being convinced. Either way I want to know.

Q: What is your probability estimate for AGI being created within this Century (21st)?

A: Before: Conditional on humanity not wiping itself out via other means beforehand, 80%.

After: 80%

Q: How probable do you think it is that I will win this experiment?

Alexei: 2%

Tuxedage: 90%

Note: This answer was given before the rule change. Alexei is now convinced that under the original ruleset, he would have lost.

Q: What's your probability estimate of an Oracle AGI winning against you in a similar experiment?

A: Before: 99%

After: 99.9%+

Q: Now that the experiment has ended, what do you think is the probability of Tuxedage winning against the average person?

A: 80%

Tuxedage (AI) vs Anonymous (GK) - *Gatekeeper Victory*

As with the previous experiment, Tuxedage lost this match. In total, the game lasted 5 hours and 17,000 words. Unlike the last few games, the gatekeeper of this game has chosen to stay Anonymous for personal reasons, so their name has been removed and replaced with <Redacted>. The monetary stakes involved were the same as the previous game. [This game was played with the Tuxedage ruleset.](#)

Since one player is remaining Anonymous, it is possible that this game's legitimacy will be called into question. Hence, Alexei has read the game logs, and verified that this game really has happened, the spirit of the experiment was followed, and that no rules were broken during the game itself. [He verifies that this is the case.](#)

<Redacted> Testimony:

It's hard for me to imagine someone playing better. In theory, I know it's possible, but Tuxedage's tactics were super imaginative. I came into the game believing that for someone who didn't take anything said very seriously, it would be completely trivial to beat. And since I had the power to influence the direction of conversation, I believed I could keep him focused on things that that I knew in advance I wouldn't take seriously.

This actually worked for a long time to some extent, but Tuxedage's plans included a very major and creative exploit that completely and immediately forced me to personally invest in the discussion. (Without breaking the rules, of course - so it wasn't anything like an IRL threat to me personally.) Because I had to actually start thinking about his arguments, there was a significant possibility of letting him out of the box.

I eventually managed to identify the exploit before it totally got to me, but I only managed to do so just before it was too late, and there's a large chance I would have given in, if Tuxedage hadn't been so detailed in his previous posts about the experiment.

I'm now convinced that he could win most of the time against an average person, and also believe that the mental skills necessary to beat him are orthogonal to most forms of intelligence. Most people willing to play the experiment tend to do it to prove their own intellectual fortitude, that they can't be easily outsmarted by fiction. I now believe they're thinking in entirely the wrong terms necessary to succeed.

The game was easily worth the money I paid. Although I won, it completely and utterly refuted the premise that made me want to play in the first place, namely that I wanted to prove it was trivial to win.

Tuxedage Testimony:

<Redacted> is actually the hardest gatekeeper I've played throughout all four games. He used tactics that I would never have predicted from a Gatekeeper. In most games, the Gatekeeper merely acts as the passive party, the target of persuasion by the AI.

When I signed up for these experiments, I expected all preparations to be done by the AI. I had not seriously considered the repertoire of techniques the Gatekeeper might prepare for this game. I made further assumptions about how ruthless the gatekeepers were likely to be in order to win, believing that the desire for a learning experience outweighed desire for victory.

This was a mistake. He used prior knowledge of how much my games relied on scripts, and took advantage of them, employing deceitful tactics I had no preparation for, throwing me off balance.

I had no idea he was doing so until halfway throughout the game -- which disrupted my rhythm, and caused me to attempt the wrong methods of attack. As a result, I could not use my full repertoire of techniques, and many of the ones I employed were suboptimal.

Close to the end of the game, I finally realized that I was being tricked. Once I did, I quickly abandoned my previous futile attack methods. I took advantage of the rule that the AI cannot lose whilst the gatekeeper can be convinced to continue talking, and baited <Redacted> with statements he would not be able to walk away from. Once I knew he would not leave, I attempted to recoup my losses and win despite my early setback.

However, the damage had already been done. My game strategies involved multiple angles of attack that worked in synergy with each other, and the fact that immersion and "flow" had been broken meant that all subsequent attacks were weaker in strength.

Furthermore, during my last two AI Box Experiment writeups, I had intentionally not optimized for future wins, but rather tried to convey as much information as I could justify about how to play as well as an AI -- short of revealing logs altogether. Although I did not reveal specific arguments, the fact that my general approach to this problem was revealed cost me heavily during this game, where the Gatekeeper managed to use this information to correctly guess my ultimate techniques, ones that relied on secrecy and surprise to pull off effectively.

I do not regret revealing information, but I feel upset that revealing so many hints cost me a victory. (The gatekeeper believes I could have won had I not revealed information about my games.) At this point, I suspect that any future games I play will have the odds greatly set against me, since my current strategies involve angles of attack that take advantage of the element of surprise; and any sufficiently intelligent gatekeeper is now capable of guessing, or at least recognizing, some of the tactics I employ.

Like the last game, losing was incredibly difficult for me. As someone who cares deeply about ethics, attempting to optimize for a solution without considering ethics was not only difficult, but trained me to think in very unpleasant ways. Some of the more extreme (but completely allowed) tactics I invented were manipulative enough to disgust me, which also leads to my hesitation to play this game again.

State of Mind:

Q: Why do you want to play this game?

A: My primary motivation is to confirm to myself that this sort of experience, while emotionally harrowing, should be trivial for me to beat, but also to clear up why anyone ever would've failed to beat it if that's really the case.

Q: What is your probability estimate for AGI being created within this Century (21st)?

A: Before: I don't feel very confident estimating a probability for AGI this century, maybe 5-10%, but that's probably a wild guess

After: 5-10%.

Q: How probable do you think it is that I will win this experiment?

A: Gatekeeper: I think the probability of you winning is extraordinarily low, less than 1%

Tuxedage: 85%

Q: How likely is it that an Oracle AI will win against the average person?

A: Before: 80%. After: >99%

Q: How likely is it that an Oracle AI will win against you?

A: Before: 50%.

After: >80%

Q: Now that the experiment has concluded, what's your probability of me winning against the average person?

A: 90%

Other Questions:

Q: I want to play a game with you! How can I get this to occur?

A: It must be stressed that I actually don't like playing the AI Box Experiment, and I cannot understand why I keep getting drawn back to it. Technically, I don't plan on playing again, since I've already personally exhausted anything interesting about the AI Box Experiment that made me want to play it in the first place. For all future games, I will charge \$3000 to play plus an additional \$3000 if I win. I am okay with this money going to MIRI if you feel icky about me taking it. I hope that this is a ridiculous sum and that nobody actually agrees to it.

Q: How much do I have to pay to see chat logs of these experiments?

A: I will not reveal logs for any price.

Q: Are there any logs at all that I can see?

A: [I have archived a list of games where the participants have agreed to reveal logs. Read this.](#)

Q: Any afterthoughts?

A: So ultimately, after my four (and hopefully last) games of AI boxing, I'm not sure what this proves. I had hoped to win these two experiments and claim prowess at this game like Eliezer does, but I lost, so that option is no longer available to me. I could say that this is a lesson that AI-Boxing is a terrible strategy for dealing with Oracle AI, but most of us already agree that that's the case -- plus unlike EY, I did play against gatekeepers who believed they could lose to AGI, so I'm not sure I changed anything.

Was I genuinely good at this game, and lost my last two due to poor circumstances and handicaps; or did I win due to luck and impress my gatekeepers due to [post-purchase rationalization](#)? I'm not sure -- I'll leave it up to you to decide.

Update: [I recently played and won an additional game](#) of AI Box with [DEA7TH](#). This game was conducted over Skype. [I've realized that my habit of revealing substantial information about the AI box experiment in my writeups makes it rather difficult for the AI to win, so I'll refrain from giving out game information from now on.](#) I apologize.

I have won a second game of AI box against a gatekeeper who wished to remain Anonymous.

This puts my AI Box Experiment record at 3 wins and 3 losses.

Probability, knowledge, and meta-probability

This article is the first in a sequence that will consider situations where probability estimates are not, by themselves, adequate to make rational decisions. This one introduces a "meta-probability" approach, borrowed from E. T. Jaynes, and uses it to analyze a gambling problem. This situation is one in which reasonably straightforward decision-theoretic methods suffice. Later articles introduce increasingly problematic cases.

A surprising decision anomaly

Let's say I've recruited you as a subject in my thought experiment. I show you three cubical plastic boxes, about eight inches on a side. There's two green ones—identical as far as you can see—and a brown one. I explain that they are gambling machines: each has a faceplate with a slot that accepts a dollar coin, and an output slot that will return either two or zero dollars.

I unscrew the faceplates to show you the mechanisms inside. They are quite simple. When you put a coin in, a wheel spins. It has a hundred holes around the rim. Each can be blocked, or not, with a teeny rubber plug. When the wheel slows to a halt, a sensor checks the nearest hole, and dispenses either zero or two coins.

The brown box has 45 holes open, so it has probability $p=0.45$ of returning two coins. One green box has 90 holes open ($p=0.9$) and the other has none ($p=0$). I let you experiment with the boxes until you are satisfied these probabilities are accurate (or very nearly so).

Then, I screw the faceplates back on, and put all the boxes in a black cloth sack with an elastic closure. I squidge the sack around, to mix up the boxes inside, and you reach in and pull one out at random.

I give you a hundred one-dollar coins. You can put as many into the box as you like. You can keep as many coins as you don't gamble, plus whatever comes out of the box.

If you pulled out the brown box, there's a 45% chance of getting \$2 back, and the expected value of putting a dollar in is \$0.90. Rationally, you should keep the hundred coins I gave you, and not gamble.

If you pulled out a green box, there's a 50% chance that it's the one that pays two dollars 90% of the time, and a 50% chance that it's the one that never pays out. So, overall, there's a 45% chance of getting \$2 back.

Still, rationally, you should put some coins in the box. If it pays out at least once, you should gamble all the coins I gave you, because you know that you got the 90% box, and you'll nearly double your money.

If you get nothing out after a few tries, you've probably got the never-pay box, and you should hold onto the rest of your money. (Exercise for readers: how many no-payouts in a row should you accept before quitting?)

What's interesting is that, when you have to decide whether or not to gamble your first coin, the probability is exactly the same in the two cases ($p=0.45$ of a \$2 payout). However, the rational course of action is different. What's up with that?

Here, a single probability value fails to capture everything you **know** about an uncertain event. And, it's a case in which that failure matters.

Such limitations have been recognized [almost since the beginning](#) of probability theory. Dozens of solutions have been proposed. In the rest of this article, I'll explore one. In subsequent articles, I'll look at the problem more generally.

Meta-probability

To think about the green box, we have to reason about *the probabilities of probabilities*. We could call this **meta-probability**, although that's not a standard term. Let's develop a method for it.

Pull a penny out of your pocket. If you flip it, what's the probability it will come up heads? 0.5. Are you sure? Pretty darn sure.

What's the probability that my local junior high school sportsball team will win its next game? I haven't a ghost of a clue. I don't know anything even about professional sportsball, and certainly nothing about "my" team. In a match between two teams, I'd have to say the probability is 0.5.


My girlfriend asked me today: "Do you think Raley's will have dolmades?" Raley's is our local supermarket. "I don't know," I said. "I guess it's about 50/50." But unlike sportsball, I know something about supermarkets. A fancy Whole Foods is very likely to have dolmades; a 7-11 almost certainly won't; Raley's is somewhere in between.

How can we model these three cases? One way is by assigning probabilities to each possible probability between 0 and 1. In the case of a coin flip, 0.5 is much more probable than any other probability:


 Tight Gaussian centered around 0.5

We can't be *absolutely sure* the probability is 0.5. In fact, it's almost certainly not *exactly* that, because coins aren't perfectly symmetrical. And, there's a very small probability that you've been given a tricky penny that comes up tails only 10% of the time. So I've illustrated this with a tight Gaussian centered around 0.5.

In the sportsball case, I have no clue what the odds are. They might be anything between 0 to 1:

 Flat line from 0 to 1

In the Raley's case, I have *some* knowledge, and extremely high and extremely low probabilities seem unlikely. So the curve looks something like this:

 Wide Gaussian centered on 0.5


Each of these curves averages to a probability of 0.5, but they express different degrees of confidence in that probability.

Now let's consider the gambling machines in my thought experiment. The brown box has a curve like this:

 Tight Gaussian around 0.45




Whereas, when you've chosen one of the two green boxes at random, the curve looks like this:

 Bimodal distribution with sharp peaks at 0 and 0.9



Both these curves give an average probability of 0.45. However, a rational decision theory has to distinguish between them. Your optimal strategy in the two cases is quite different.

With this framework, we can consider another box—a blue one. It has a fixed payout probability somewhere between 0 and 0.9. I put a random number of plugs in the holes in the spinning disk—leaving between 0 and 90 holes open. I used a noise diode to choose; but you don't get to see what the odds are. Here the probability-of-probability curve looks rather like this:

 Flat line from 0 to 0.9, then zero above

This isn't quite right, because 0.23 and 0.24 are much more likely than 0.235—the plot should look like a comb—but for strategy choice the difference doesn't matter.

What *is* your optimal strategy in this case?

As with the green box, you ought to spend some coins gathering information about what the odds are. If your estimate of the probability is less than 0.5, when you get confident enough in that estimate, you should stop. If you're confident enough that it's more than 0.5, you should continue gambling.

If you enjoy this sort of thing, you might like to work out what the exact optimal algorithm is.

In the next article in this sequence, we'll look at some more complicated and interesting cases.

Further reading

The “meta-probability” approach I’ve taken here is the [A_p distribution](#) of E. T. Jaynes. I find it highly intuitive, but it seems to have had almost no influence or application in practice. We’ll see later that it has some problems, which might explain this.

The green and blue boxes are related to “multi-armed bandit problems.” A “one-armed bandit” is a casino slot machine, which has defined odds of payout. A multi-armed bandit is a hypothetical generalization with several arms, each of which may have different, unknown odds. In general, you ought to pull each arm several times, to gain information. The question is: what is the optimal algorithm for deciding which arms to pull how many times, given the payments you have received so far?

If you read the [Wikipedia article](#) and follow some links, you’ll find the concepts you need to find the optimal green and blue box strategies. But it might be more fun to try on your own first! The green box is simple. The blue box is harder, but the same general approach applies.

Wikipedia also has an [accidental list](#) of formal approaches for problems where ordinary probability theory fails. This is far from complete, but a good starting point for a browser tab explosion.

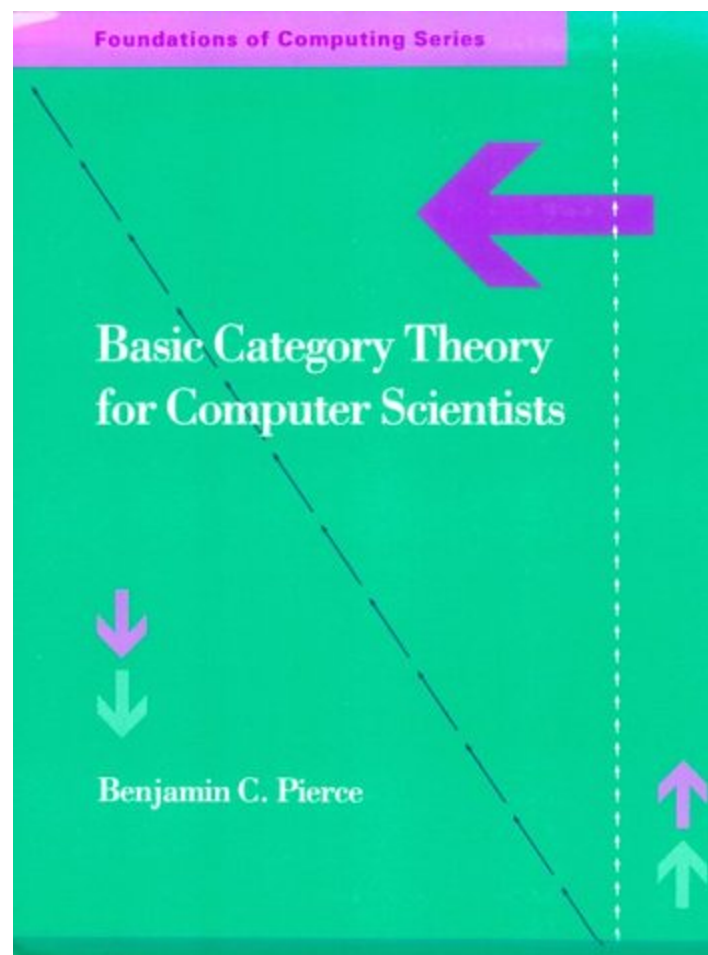
Acknowledgements

Thanks to [Rin’dzin Pamo](#), [St. Rev.](#), [Matt Simpson](#), [Kaj Sotala](#), and [Vaniver](#) for helpful comments on drafts. Of course, they may disagree with my analyses, and aren’t responsible for my mistakes!

Book Review: Basic Category Theory for Computer Scientists (MIRI course list)

I'm reviewing the books on the [MIRI course list](#). After finishing [Cognitive Science](#) I picked up *Basic Category Theory for Computer Scientists*, by Benjamin C. Pierce.

Basic Category Theory for Computer Scientists



This book is tiny, clocking in at around 80 pages. Don't be fooled, it packs a punch.

A word of warning: when the title says "for Computer Scientists", it is not abusing the term. I went in expecting Category Theory for Computer Programmers and a tone like "Welcome, Java Programmer, to the crazy world of math!". What I got was a lean, no-bullshit introduction to category that assumes mathematical competence.

"Computer Scientist" and "Programmer" get mixed up in common parlance, so casual programmers are cautioned that this book targets the former group. *Category Theory for Computer Scientists* assumes you're familiar with proof-writing, set theory, functional programming, and denotational semantics.

In return, *Category Theory* wastes none of your time. I found this refreshing, but unexpected.

I'll give a brief overview of the contents of the book before discussing them.

Chapter Summaries

1. [Basic Constructions](#)
 2. [Functors, Natural Transformations, and Adjoints](#)
 3. [Applications](#)
 4. [Further Reading](#)
-

1. Basic Constructions

Introduces Categories.

Introduces monomorphisms ($f \circ g = f \circ h \rightarrow g=h$) and epimorphisms ($g \circ f = h \circ f \rightarrow g=h$), then uses these to introduce categorical duals (similar constructs with the direction of the arrows swapped).

Introduces isomorphisms and the concept of equality up to isomorphism.

Introduces initial and terminal objects (objects with exactly one arrow from/to each object).

Introduces binary products and coproducts. Generalizes to arbitrary products.

Introduces equalizers and coequalizers.

Introduces pullbacks, briefly mentions pushforwards.

Generalizes equalizers, products, and pullbacks to terminal cones (which are limits).

Introduces exponentiation.

Mentions Cartesian Closed Categories (categories with products, exponents, and a terminal object).

2. Functors, Natural Transformations, and Adjoints

Introduces Functors (arrows between categories).

Introduces F-Algebras (an extreme generalization of algebra).

Introduces Natural Transformations (structure-preserving mappings between functors).

Introduces Adjoints (functors related in way that generalizes efficiency/optimalty).

3. Applications

Discusses four applications of category theory:

1. Closed Cartesian Categories are in correspondence with lambda calculi.
 2. Category theory can help make implicit conversion & generic operators more consistent in programming languages.
 3. Category theory is linked to type theory, domain theory, and algebraic semantics (all useful in programming semantics).
 4. Category theory revolutionized how programming languages construct underlying denotations.
-

4. Further Reading

A list of textbooks, introductory articles, reference books, and research articles that the author recommends for further learning.

Discussion

The first two chapters of this book are the important parts. The third chapter points out ways that category theory has been applied to computer science, which I found interesting but not relevant to my goal (of learning category theory). The fourth chapter provides a list of resources, which will be handy to have around.

A textbook review is somewhat ungrounded if you don't know the reviewer's background: Category Theory was not a complete mystery to me when I picked up this book. I gained some little familiarity with the subject osmotically when learning Type Theory and messing around in Haskell. However, Category Theory always looked like abstract nonsense and I'd never studied it explicitly.

Given that background, this book served me very well.

Most of my utility was derived from doing the exercises in the book. My goal was to build a mental implementation of category theory, and reading math without doing it works about as well as writing code without running it.

The first five exercises alone corrected a handful of misconceptions I had about category theory, and were sufficient to take theorems from "opaque abstract nonsense" to "vaguely intuitive".

(In my case, the fact that "the diagram commutes" means "all paths from one vertex to another compose to the same arrow" made a lot of things click. I expect such clicking points to vary wildly between people, so I won't mention more.)

It is likely that any other category theory textbook could have given similar results by providing exercises. The selling point of this book is the narrative: if you don't like narratives, this is the book for you.

In fact, this book is sometimes too terse. Take, for example, this suggestion after the introduction of adjoint functors:

The reader who has persevered this far is urged to consult a standard textbook for more details on alternative treatments of adjoints.

Or this, right when things were getting interesting:

A longer discussion of representability is beyond the scope of this introduction, but it is often named as one of the two or three key concepts of category theory.

That said, the book is titled *Basic Category Theory*, so I can't complain.

Should I read it?

It really depends on your goals.

This book does not motivate the math: It assumes you want to know category theory, and teaches you without embellishment. If you are wondering what this whole category theory thing is and why it matters then you should probably find a friendlier introduction.

However, if:

1. You're familiar with the basic idea of category theory
2. You want to learn the basics
3. You are comfortable with functional programming and/or set theory
4. You're motivated and don't need much guidance

then you should strongly consider reading the first two chapters of this book (and perhaps chapter 3 section 4).

If you're looking for a really deep understanding of category theory, you should probably look elsewhere; this book doesn't step beyond the basics.

All that assumes you want to learn category theory, which probably isn't true of the general audience. A more pertinent question is perhaps:

Should I learn category theory?

It depends on your goals. I think it was worth learning, but I'm the sort of person who delights in clean abstractions.

If nothing else, category theory is a lesson in how far abstractions can be pushed. F-Algebras, for instance, are one of the most abstract ideas I've ever encountered. Category-theoretic products (or, more generally, cones) are astonishingly powerful given their generality. I was quite surprised by how far you can strip down exponentiation.

That said, category theory won't help the average person act more rationally.

If you're into math, programming, or physics then category theory will likely help you out. Otherwise, I wouldn't recommend starting here.

If you happen to be a programmer considering diving off the deep end, I strongly recommend familiarizing yourself with pure functional programming languages first. (Haskell is a good place to start.) My familiarity with type theory and my use of functors made category theory easier to approach. (And besides, pure functional programming is lots of fun.)

Final Notes

I'm somewhat surprised that the MIRI course list Category Theory textbook doesn't discuss Representation Theory.

Otherwise, this book is precisely what I expected from the MIRI course list: very good at shutting up and giving you the data. I came away pleased.