Best of LessWrong: January 2012

1. How I Ended Up Non-Ambitious
2. Can the Chain Still Hold You?
3. [Transcript] Richard Feynman on Why Questions
4. Urges vs. Goals: The analogy to anticipation and belief
5. The problem with too many rational memes
6. The Substitution Principle
7. The Singularity Institute's Arrogance Problem
8. Completeness, incompleteness, and what it all means: first versus second order logic
9. So You Want to Save the World
10. The Noddy problem
11. [META] My Negative Results
12. The Human's Hidden Utility Function (Maybe)
13. Introducing Leverage Research
14. Shit Rationalists Say?
15. "Politics is the mind-killer" is the mind-killer
16. Leveling Up in Rationality: A Personal Journey
17. What Curiosity Looks Like
18. The Personality of (great/creative) Scientists: Open and Conscientious
19. [META] 'Rational' vs 'Optimized'

Best of LessWrong: January 2012

How I Ended Up Non-Ambitious

I have a confession to make. My life hasn't changed all that much since I started reading Less Wrong. Hindsight bias makes it hard to tell, I guess, but I feel like pretty much the same person, or at least the person I would have evolved towards anyway, whether or not I spent those years reading about the Art of rationality.

But I can't claim to be upset about it either. I can't say that rationality has undershot my expectations. I didn't come to Less Wrong expecting, or even wanting, to become the next Bill Gates; I came because I enjoyed reading it, just like I've enjoyed reading hundreds of books and websites.

In fact, I can't claim that I would *want* my life to be any different. I have goals and I'm meeting them: my grades are good, my social skills are slowly but steadily improving, I get along well with my family, my friends, and my boyfriend. I'm in good shape financially despite making $12 an hour as a lifeguard, and in a year and a half I'll be making over $50,000 a year as a registered nurse. I write stories, I sing in church, I teach kids how to swim. Compared to many people my age, I'm pretty successful. In general, I'm pretty happy.

Yvain suggested akrasia as a major limiting factor for why rationalists fail to have extraordinarily successful lives. Maybe that's true for some people; maybe they are some readers and posters on LW who have big, exciting, challenging goals that they consistently fail to reach because they lack motivation and procrastinate. But that isn't true for me. Though I can't claim to be totally free of akrasia, it hasn't gotten much in the way of my goals.

However, there are some assumptions that go too deep to be accessed by introspection, or even by LW meetup discussions. Sometimes you don't even realize they're assumptions until you meet someone who assumes the opposite, and try to figure out why they make you so defensive. At the community meetup I described in my last post, a number of people asked me why I wasn't studying physics, since I was obviously passionate about it. Trust me, I had plenty of good justifications for them– it's a question I've been asked many times–but the question itself shouldn't have made me feel attacked, and it did.

Aside from people in my life, there are some posts on Less Wrong that cause the same reaction of defensiveness. Eliezer's Mandatory Secret Identities is a good example; my automatic reaction was "well, why do you assume everyone here wants to have a super cool, interesting life? In fact, why do you assume everyone wants to be a rationality instructor? I don't. I want to be a nurse."

After a bit of thought, I've concluded that there's a simple reason why I've achieved all my life goals so far (and why learning about rationality failed to affect my achievements): they're not hard goals. I'm not ambitious. As far as I can tell, not being ambitious is such a deep part of my identity that I never even noticed it, though I've used the underlying assumptions as arguments for why my goals and life decisions were the right ones.

But if there's one thing Less Wrong has taught me, it's that assumptions are to be questioned. There are plenty of good reasons to choose reasonable goals instead of impossible ones, but doing things on reflex is rarely better than thinking through

them, especially for long-term goal making, where I do have time to think it through, Type 2 style.

What do I mean by ‘ambition’?

Here is the definition from my desktop dictionary:

(1) A strong desire to do or to achieve something, typically requiring determination and hard work*: her ambition was to become a model | he achieved his ambition of making a fortune.*

(2) Desire and determination to achieve success*: life offered few opportunities for young people with ambition.*

The first definition sounds like a good description of me. Since around tenth grade, I’ve had a strong desire to study nursing, and it’s required a moderate amount of determination and hard work, especially the hands-on aspects, which are harder for me than academics has ever been. I want to be the kind of person described in (1).

What about the second half? More people than I can count have asked me why I’m not studying medicine. Or physics. Or just about anything aside from nursing, which is apparently kind of low-status. I inevitably get defensive when these conversations occur, and I end up trying to justify why nursing is the morally correct thing for me to do. For some reason, in some deep-down part of me that I don’t normally have conscious access to, I don’t want to be the sort of person described in (2).

Introspection isn’t accurate enough for me to automatically find my true rejection of ambitious goals, but I will take the rest of the post to speculate on *my own personal* reasons. They may or may not be reasons that generalize to anyone else.

1. Idealism versus practicality

My mother tells me I would be a good academic, and enjoy it too. She’s usually right about that kind of thing, but I decided around eighth grade that academia wasn’t for me.

Why? Well, my mother and father both studied science at the undergraduate level (biology and physical chemistry, respectively) and then both went on to complete PhDs. From the sound of it, those student years were among the happiest in their lives. My father went on to do a postdoc at Cambridge, and then to get a crappy part-time teaching position at a small university in Washington State. He hated it. Eventually he quit and we moved up to Ottawa, Canada, where he worked at Nortel, was laid off during the company’s decline, and eventually found another job at a small company that takes apart computer chips and analyzes them. Meanwhile, my mother spent most of those years as a housewife, and has only recently begun working again, part-time and for a token salary.

I've asked my father what he thinks of the decisions he made, and he told me that his biggest problem was that he didn't know what he wanted to do with his life. He told me that he still doesn't. His job is boring and stressful, but he can't quit because he didn't start saving for retirement until he was 40. As a grad student, he worked with John Polanyi, a well-known academic; much later he told me he "always sort of thought I would end up being well-known and cool like that, but all of a sudden I'm almost 50 and I realize that's not going to happen."

I remember the year when he developed a sudden passion for career self-help books, of the 'What Color Is Your Parachute' and 'The Seven Habits of Highly Effective People' variety. I must have been about thirteen years old. He encouraged me to read them, and warned me that "it's better to think about what you want to do, not what you want to be."

The lesson my 13-year-old self I took from all this: don't have hopes and dreams, especially not ambitious ones. You won't achieve them, and you'll end up in a mid-life crisis with no retirement savings, full of regrets. Far better to have a practical, achievable life plan, and then go out and damn well *achieve* it. I read the self-help books, figured that nurses did around the same stuff all day as doctors and didn't have to spend eight years in school paying tuition, and never looked back.

The lesson I *didn't* learn from all this: my parents weren't actually ambitious either. They enjoyed their studies in university, but primarily they had fun: going to the philosophy faculty parties, getting drunk with chemistry students, volunteering on coffee plantations in Nicaragua... Those are the stories they tell me from their studies, not stories of the research they did and the papers they published. I can't be sure what their true feelings were at the time, but I don't think they *cared* especially. They were smart young people who wanted to have a good time and didn't especially care if they had no money. And I don't think they have as many regrets as I assumed when I was thirteen. They didn't exactly make life goals and then fail to achieve them. They just hadn't made their long-term goals ahead of time.

The lesson I *should* have learned: if you head into adulthood without big goals, don't be surprised if you don't achieve them.

2. Fear of failure

The second life lesson about ambition happened a few years later, when I was around fourteen. I had been training as a competitive swimmer for a number of years. My parents didn't sign me up because they wanted me to go to the Olympics someday; they wanted me to stay fit and have opportunities to socialize. It was a good decision; swim team made me happy, to the point that I often forget how unhappy I was up until then.

But after a while I started to get *good* at swimming, and coaches, even kids' coaches, implicitly want their athletes to win, and keep winning, and maybe someday they'll be known as the one who coached an Olympic athlete. Training made me happy, but competition emphatically did *not*; anxiety, stress, and bursting into tears before a race soon became part of my day-to-day life. My coaches told me that if I worked hard and believed in myself, I could do anything. But eventually I hit a point when I was racing kids who were simply *more talented* than me: taller, slimmer, bigger hands and feet, a

genetic predisposition to fast-twitch muscles, whatever. And then I hit my body's limits, and I stopped getting faster at all, no matter how hard I trained. My coaches accused me of not trying hard enough. Understandably, this made me feel worse, since I certainly *felt* like I was trying as hard as I could.

The lessons my 14-year-old self learned from this: don't have high expectations for yourself when competing against other people. You'll just end up feeling worthless and depressed. In fact, don't compete against other people *at all*. Do things that are solely based on how good you are, as opposed to how good you are *relative to other people who might be more talented.* Even better, do things that aren't that hard in an absolute sense, so that you don't risk failing.

This is kind of a fallacy, of course. Success in anything is measured relative to other people, if only relative to the average. Even grades, because classes and tests and grades are set up for students of average intelligence, so students of relatively higher intelligence will find them easier, and students of lower-than-average intelligence will feel like they're fighting a losing battle, as I did in swimming competitions. Possessing above average intelligence let me grow up seeing school as non-threatening, but I know that isn't true for everyone. I've known people whose above-average athletic skills led them to be far more confident in sports than at school.

Still, fallacy or not, I later applied this idea to a lot of my decision. I was interested in physics all along, but my father's tales of academia and the competition and pressure involved turned me off it. I also considered studying music theory and composition, but decided not to because, aside from being impractical for finding a job afterwards, I'd heard it was an incredibly competitive field. To a degree, this is why I chose not to make a career as a writer. (A degree in English didn't seem particularly interesting to me, so I doubt I would have studied it, but even in high school I never really thought about earning money with my writing.) Success or failure was too far beyond my control for comfort.

The lesson I didn't learn from this: find an area where you *do* have natural talent on your side, and use it for all it's worth. In fact, I've done the opposite of this: one reason I chose nursing was because I felt that I was bad at a whole range of skills; empathy, social skills, fine motor skills and coordination, reacting in emergencies; and I wanted to force myself to improve. As a result, I'm far from the strongest student in my classes, and labs, simulations, and hospital placements bring me to a level of anxiety far above anything I ever experienced during academic tests or exams.

The lesson I should have learned from this: you never know what you are and aren't capable of until you try it. I tried competitive swimming, and found out I didn't have the raw talent to go to the Olympics. Who knows if this would have been true of physics? My father tells me that in his fourth year of undergraduate studies, he took several physics courses with a level of advanced math that he found almost impossible. He had reached his brain's natural limit in math, which he might or might not have been able to exceed with hard work and hours of study; still, it was *much* more advanced than the first-year calculus I took as an elective. I have no reason to think that I'm *worse* at math than my father, and I suspect my obsessive work ethic would help me exceed any limits I did bump up against. And why not try?

3. The morality of ambition

There’s a third aspect of my aversion to ambitious goals, and I can’t say where it comes from. It might be my parents’ attitude of moderation in everything: they consistently disapproved of my involvement in any ‘obsessive’ activities, swim team included. It might be the way my mother always got mad at me for talking about my achievements, even my grades, in front of friends; it’ll make other people feel bad, she said. (For a long time I was incredibly self-conscious about high grades, and wouldn’t tell my friends if they were above 90%.) It might be the meme that ‘money doesn’t buy happiness’ or the idea that it’s greedy to be ambitious, or that power corrupts and wise people choose not to seek it.

I can’t trace the roots of this idea completely, but for whatever reason, I spent a long time thinking that being ambitious was in some way *immoral*. That really good people lived simple, selfless lives and never tried to seek anything more. That doing something solely because you wanted more money or more respect, like going to med school instead of nursing school, was selfish and just *bad*. It might come from the books I read as a kid, or maybe it’s just a rationalization to cover up my other reasons with a nobler one.

But if this is my true reason, then it’s a way to feel superior to people who’ve accomplished cooler things than me, of whom part of me is actually jealous, and that’s *not* the person I want to be.

4. Laziness

I don’t normally think of myself as a lazy person. Other people are constantly telling me that I’m diligent and have an excellent work ethic. But there’s a way in which all this hardworking dedication to my current occupations has *prevented* me from spending much time thinking or acting about what I’m going to do next. Working a bunch of 12-hour shifts makes me feel productive, brings the direct benefit of a fat paycheck, and leaves me pretty exhausted at the end of the day, too tired to do the (in some ways harder) work of searching for cool job opportunities, looking at online classes to take, and in general breaking the routine. I *hate* breaking my routine. It makes me anxious, and I have to spend more energy motivating myself, and in general it’s *hard*. I tend to only depart from that routine when forced.

Conclusion

I think I was right about *some* of the conclusions I drew from these various experiences. Practicality is important: ask the English majors working at Starbucks. Thinking about what you want to do all day, as opposed to the title and respect associated with what you want to *be*, is good life advice and will likely result in a more satisfying career. Trying hard to project an image of success, i.e. “keeping up with the Jones’”, isn’t a good path to happiness. And relative talent is a factor to take into consideration; if my dream career were to be an Olympic swimmer, unfortunately I wouldn’t be likely to succeed.

But one of the problems with thinking things through too deeply when you’re young, and think you’re wiser than everyone else, is a tendency to over-generalize. Doing

cool, interesting, world-changing things with your life...even if the actual job position are competitive and hard to obtain...well, on reflection, it doesn't seem be a *bad* idea.

The lesson my current self has learned from this: investigate more. Spend less time on work and more time on actually planning future goals. Seek out interesting things to do, and interesting people to work with. Go for opportunities even if they're inconvenient and I have to break my routine a bit. Set *concrete* goals, and don't wiggle out of achieving them because they're 'not actually that important.' They're probably more important than working at a community centre, and I seem to be able to dedicate 1000 hours a year to that... Try not to worry about sunk costs (although it's worth finishing nursing school, since an RN certificate is incredibly versatile in Canada and will guarantee me a job if any other prospects fail.) Force myself to step out of my comfort zone once in a while and do something kind of crazy, but awesome. And if I can do that, succeed to the point that I can break my reflex-of-being-average...*then* I'll know for sure whether rationality, of the Less Wrong variety, will help me to 'win'.

The lesson my future self will learn from this: who knows?

Can the Chain Still Hold You?

Robert Sapolsky:

> Baboons... *literally* have been the textbook example of a highly aggressive, male-dominated, hierarchical society. Because these animals hunt, because they live in these aggressive troupes on the Savannah... they have a constant baseline level of aggression which inevitably spills over into their social lives.

Scientists have never observed a baboon troupe that *wasn't* highly aggressive, and they have compelling reasons to think this is simply *baboon nature*, written into their genes. Inescapable.

Or at least, that was true until the 1980s, when Kenya experienced a tourism boom.

Sapolsky was a grad student, studying his first baboon troupe. A new tourist lodge was built at the edge of the forest where his baboons lived. The owners of the lodge dug a hole behind the lodge and dumped their trash there every morning, after which the males of several baboon troupes — including Sapolsky's — would fight over this pungent bounty.

Before too long, someone noticed the baboons didn't look too good. It turned out they had eaten some infected meat and developed tuberculosis, which kills baboons in weeks. Their hands rotted away, so they hobbled around on their elbows. Half the males in Sapolsky's troupe died.

This had a surprising effect. There was now almost no violence in the troupe. Males often reciprocated when females groomed them, and males even groomed other males. To a baboonologist, this was like watching Mike Tyson suddenly stop swinging in a heavyweight fight to start nuzzling Evander Holyfield. It *never* happened.

This was interesting, but Sapolsky moved to the other side of the park and began studying other baboons. His first troupe was "scientifically ruined" by such a non-natural event. But really, he was just heartbroken. He never visited.

Six years later, Sapolsky wanted to show his girlfriend where he had studied his first troupe, and found that they were still there, and still surprisingly violence-free. This one troupe had apparently been so transformed by their unusual experience — and the continued availability of easy food — that they were now basically non-violent.

And then it hit him.

Only one of the males now in the troupe had been through the event. All the rest were new, and hadn't been raised in the tribe. The new males had come from the violent, dog-eat-dog world of normal baboon-land. But instead of coming into the new troupe and roughing everybody up as they *always* did, the new males had learned, "We don't do stuff like that here." They had unlearned their childhood culture and adapted to the new norms of the first baboon pacifists.

As it turned out, violence *wasn't* an unchanging part of baboon nature. In fact it changed rather quickly, when the right causal factor flipped, and — for this troupe and the new males coming in — it has *stayed* changed to this day.

Somehow, the violence had been largely *circumstantial*. It was just that the circumstances had always been the same.

Until they weren't.

We still don't know how much baboon violence to attribute to nature vs. nurture, or exactly how this change happened. But it's worth noting that changes like this can and do happen pretty often.

Slavery was ubiquitous for millennia. Until it was outlawed in every country on Earth.

Humans had never left the Earth. Until we achieved the first manned orbit and the first manned moon landing in a single decade.

Smallpox occasionally decimated human populations for thousands of years. Until it was eradicated.

The human species was always too weak to render itself extinct. Until we discovered the nuclear chain reaction and manufactured thousands of atomic bombs.

Religion had a grip on 99.5% or more of humanity until 1900, and then the rate of religious adherence plummeted to 85% by the end of the century. Whole nations became mostly atheistic, largely because for the first time the state provided people some basic stability and security. (Some nations became atheistic because of atheistic dictators, others because they provided security and stability to their citizens.)

I would never have imagined I could have the kinds of conversations I now regularly have at the Singularity Institute, where people change their degrees of belief several times in a single conversation as new evidence and argument is presented, where everyone at the table knows and applies a broad and deep scientific understanding, where people disagree strongly and say harsh-sounding things (due to Crocker's rules) but end up coming to agreement after 10 minutes of argument and carry on as if this is friendship and business as usual — because it is.

But then, never before has humanity had the combined benefits of an overwhelming case for one correct probability theory, a systematic understanding of human biases and how they work, free access to most scientific knowledge, and a large community of people dedicated to the daily practice of CogSci-informed rationality exercises and to helping each other improve.

This is part of what gives me a sense that more is possible. Compared to situational effects, we tend to overestimate the effects of lasting dispositions on people's behavior — the fundamental attribution error. But I, for one, was only taught to watch out for this error in explaining the behavior of *individual* humans, even though the bias also appears when explaining the behavior of humans *as a species*. I suspect this is *partly* due to the common misunderstanding that heritability measures the degree to which a trait is due to genetic factors. Another reason may be that for obvious reasons scientists rarely try *very* hard to measure the effects of exposing human subjects to radically different environments like an artificial prison or total human isolation.

When taming a baby elephant, its trainer will chain one of its legs to a post. When the elephant tries to run away, the chain and the post are strong enough to keep it in place. But when the elephant grows up, it is strong enough to break the chain or uproot the post. Yet the owner can still secure the elephant with the same chain and

post, because the elephant has been conditioned to believe it cannot break free. It feels the tug of the chain and gives up — a kind of learned helplessness. The elephant acts as if it thinks the chain's limiting power is intrinsic to nature rather than dependent on a causal factor that held for years but holds no longer.

Much has changed in the past few decades, and much will change in the coming years. Sometimes it's good to check if the chain can still hold you. Do not be tamed by the tug of history. Maybe with a few new tools and techniques you can just get up and walk away — to a place you've never seen before.

[Transcript] Richard Feynman on Why Questions

I thought this video was a *really* good question dissolving by Richard Feynman. But it's in 240p! Nobody likes watching 240p videos. So I transcribed it. (*Edit*: That was in jest. The real reasons are because I thought I could get more exposure this way, and because a lot of people appreciate transcripts. Also, Paul Graham speculates that the written word is universally superior than the spoken word for the purpose of ideas.) I was going to post it as a rationality quote, but the transcript was sufficiently long that I think it warrants a discussion post instead.

Here you go:

> Interviewer: If you get hold of two magnets, and you push them, you can feel this pushing between them. Turn them around the other way, and they slam together. Now, what is it, the feeling between those two magnets?
>
> Feynman: What do you mean, "What's the feeling between the two magnets?"
>
> Interviewer: There's something there, isn't there? The sensation is that there's something there when you push these two magnets together.
>
> Feynman: Listen to my question. What is the meaning when you say that there's a feeling? Of course you feel it. Now what do you want to know?
>
> Interviewer: What I want to know is what's going on between these two bits of metal?
>
> Feynman: They repel each other.
>
> Interviewer: What does that mean, or why are they doing that, or how are they doing that? I think that's a perfectly reasonable question.
>
> Feynman: Of course, it's an excellent question. But the problem, you see, when you ask *why* something happens, how does a person answer why something happens? For example, Aunt Minnie is in the hospital. Why? Because she went out, slipped on the ice, and broke her hip. That satisfies people. It satisfies, but it wouldn't satisfy someone who came from another planet and who knew nothing about why when you break your hip do you go to the hospital. How do you get to the hospital when the hip is broken? Well, because her husband, seeing that her hip was broken, called the hospital up and sent somebody to get her. All that is understood by people. And when you explain a *why*, you have to be in some framework that you allow something to be true. Otherwise, you're perpetually asking why. Why did the husband call up the hospital? Because the husband is interested in his wife's welfare. Not always, some husbands aren't interested in their wives' welfare when they're drunk, and they're angry.
>
> And you begin to get a very interesting understanding of the world and all its complications. If you try to follow anything up, you go deeper and deeper in various directions. For example, if you go, "Why did she slip on the ice?" Well, ice is slippery. Everybody knows that, no problem. But you ask *why is ice slippery?* That's kinda curious. Ice is extremely slippery. It's very interesting. You say, how

does it work? You could either say, "I'm satisfied that you've answered me. Ice is slippery; that explains it," or you could go on and say, "Why is ice slippery?" and then you're involved with something, because there aren't many things as slippery as ice. It's very hard to get greasy stuff, but that's sort of wet and slimy. But a solid that's so slippery? Because it is, in the case of ice, when you stand on it (they say) momentarily the pressure melts the ice a little bit so you get a sort of instantaneous water surface on which you're slipping. Why on ice and not on other things? Because water expands when it freezes, so the pressure tries to undo the expansion and melts it. It's capable of melting, but other substances get cracked when they're freezing, and when you push them they're satisfied to be solid.

Why does water expand when it freezes and other substances don't? I'm not answering your question, but I'm telling you how difficult the *why* question is. You have to know what it is that you're permitted to understand and allow to be understood and known, and what it is you're not. You'll notice, in this example, that the more I ask why, the deeper a thing is, the more interesting it gets. We could even go further and say, "Why did she fall down when she slipped?" It has to do with gravity, involves all the planets and everything else. Nevermind! It goes on and on. And when you're asked, for example, why two magnets repel, there are many different levels. It depends on whether you're a student of physics, or an ordinary person who doesn't know anything. If you're somebody who doesn't know anything at all about it, all I can say is the magnetic force makes them repel, and that you're feeling that force.

You say, "That's very strange, because I don't feel kind of force like that in other circumstances." When you turn them the other way, they attract. There's a very analogous force, electrical force, which is the same kind of a question, that's also very weird. But you're not at all disturbed by the fact that when you put your hand on a chair, it pushes you back. But we found out by looking at it that that's the same force, as a matter of fact (an electrical force, not magnetic exactly, in that case). But it's the same electric repulsions that are involved in keeping your finger away from the chair because it's electrical forces in minor and microscopic details. There's other forces involved, connected to electrical forces. It turns out that the magnetic and electrical force with which I wish to explain this repulsion in the first place is what ultimately is the deeper thing that we have to start with to explain many other things that everybody would just accept. You know you can't put your hand through the chair; that's taken for granted. But that you can't put your hand through the chair, when looked at more closely, *why*, involves the same repulsive forces that appear in magnets. The situation you then have to explain is why, in magnets, it goes over a bigger distance than ordinarily. There it has to do with the fact that in iron all the electrons are spinning in the same direction, they all get lined up, and they magnify the effect of the force 'til it's large enough, at a distance, that you can feel it. But it's a force which is present all the time and very common and is a basic force of almost - I mean, I could go a little further back if I went more technical - but on an early level I've just got to tell you that's going to be one of the things you'll just have to take as an element of the world: the existence of magnetic repulsion, or electrical attraction, magnetic attraction.

I can't explain that attraction in terms of anything else that's familiar to you. For example, if we said the magnets attract like if rubber bands, I would be cheating you. Because they're not connected by rubber bands. I'd soon be in trouble. And secondly, if you were curious enough, you'd ask me why rubber bands tend to pull back together again, and I would end up explaining that in terms of electrical forces, which are the very things that I'm trying to use the rubber bands to

explain. So I have cheated very badly, you see. So I am not going to be able to give you an answer to why magnets attract each other except to tell you that they do. And to tell you that that's one of the elements in the world - there are electrical forces, magnetic forces, gravitational forces, and others, and those are some of the parts. If you were a student, I could go further. I could tell you that the magnetic forces are related to the electrical forces very intimately, that the relationship between the gravity forces and electrical forces remains unknown, and so on. But I really can't do a good job, any job, of explaining magnetic force in terms of something else you're more familiar with, because I don't understand it in terms of anything else that you're more familiar with.

Urges vs. Goals: The analogy to anticipation and belief

Partially in response to: The curse of identity

Related to: Humans are not automatically strategic, That other kind of status, Approving reinforces low-effort behaviors.

> *Joe studies long hours, and often prides himself on how driven he is to make something of himself. But in the actual moments of his studying, Joe often looks out the window, doodles, or drags his eyes over the text while his mind wanders. Someone sent him a link to which college majors lead to the greatest lifetime earnings, and he didn't get around to reading that either. Shall we say that Joe doesn't really care about making something of himself?*

The Inuit may not have 47 words for snow, but Less Wrongers do have at least two words for belief. We find it necessary to distinguish between:

- Anticipations, what we *actually expect to see happen;*
- Professed beliefs, the set of things we tell ourselves we "believe", based partly on deliberate/verbal thought.

This distinction helps explain how an atheistic rationalist can still get spooked in a haunted house; how someone can "believe" they're good at chess while avoiding games that might threaten that belief [1]; and why Eliezer had to actually crash a car before he viscerally understood what his physics books tried to tell him about stopping distance going up with the square of driving speed. (I helped Anna revise this - EY.)

A lot of our community technique goes into either (1) dealing with "beliefs" being an evolutionarily recent system, such that our "beliefs" often end up far screwier than our actual anticipations; or (2) trying to get our anticipations to align with more evidence-informed beliefs.

And analogously - this analogy is arguably obvious, but it's deep, useful, and easy to overlook in its implications - there seem to be two major kinds of wanting:

- **Urges**: concrete emotional pulls, produced in System 1's perceptual / autonomic processes
  (my urge to drink the steaming hot cocoa in front of me; my urge to avoid embarrassment by having something to add to my accomplishments log)
- **Goals**: things we tell ourselves we're aiming at, within deliberate/verbal thought and planning
  (I have a goal to exercise three times a week; I have a goal to reduce existential risk)

Implication 1: You can import a lot of technique for "checking for screwy beliefs" into "checking for screwy goals".

Urges, like anticipations, are relatively perceptual-level and automatic. They're harder to reshape and they're also harder to completely screw up. In contrast, the flexible, recent "goals" system can easily acquire goals that are wildly detached from what we actually do, wildly detached from any positive consequences, or both. Some

techniques you can port straight over from "checking for screwy beliefs" to "checking for screwy goals" include:

The fundamental:

- "What's the positive consequence?" This is the equivalent of "What's the evidence?" for beliefs. All the other cases involve not asking it, or not asking hard enough.

The Hansonian:

- *Goals as clothes / goals as tribal affiliation:* "*We* are people who have free software (/ communism / rationality / whatever) as our goal". Before you install Linux, do you think "What's the positive consequence of installing Linux?" or does it just seem like the sort of thing a free-software-supporter would do? (EY says: What *positive consequence* is achieved by marching in an Occupy Wall Street march? Can you remember anyone stating one, throughout the whole affair - "if we march, X will happen because of Y"?)
- *Goals as a signal of one's value as an ally:* Sheila insists that she wants to get a job. We inspect her situation and she's not trying very hard to get a job. But she's in debt to a lot of her friends and is borrowing more to live on a month-to-month basis. It's not hard to see why Sheila would internally profess strongly that she has a goal of getting a job.
- *Goals as personal fashion statements:* A T-Shirt that says "Give me coffee and no one gets hurt" seems to state a very strong desire for coffee. This is clearly a goal professed directly to affect how others see you, and it's more a question of affecting a 'style' than anything directly tribal or status-y.

The satiating:

- *Having goals as optimism:* "I intend to lose weight" can be created by much the same sort of internal processes that would make you believe "I will lose weight", in cases where the goal (belief) would not yet seem very plausible to an outside view.
- *Having goals as apparent progress:* My current to-do list has "write thank-you notes for wedding gifts". This makes me feel like I've appeased the demand for internal attention by having a goal. (EY: I have "send Anna and Carl their wedding gift" on my todo list. This was very effective at appeasing the need to send them a wedding gift.)

Implication 2: "Status" / "prestige" / "signaling" / "people don't really care about" is way overused to explain goal-urge delinkages that can be more simply explained by "humans are not agents".

This post was written partially in response to The Curse of Identity, wherein Kaj recounts some suboptimal goal-action linkages - wanting to contribute to the Singularity, then teaching himself to feel guilty whenever not working; founding the Finnish Pirate Party, then becoming the spokesperson which involved tasks he wasn't good at; helping Eliezer on writing his book, and feeling demotivated because it seemed like work "anyone could do" (which is just the sort of work that almost nobody is motivated to do).

Kaj forms the generalization "as soon as my brain adopted a cause, my subconscious reinterpreted it as the goal of giving the impression of doing prestigious work for the cause". I worry that our community has a tendency to explain as e.g. status

signaling or "people really don't care about X", observations that can also be explained by less malice/selfishness and more "our brains have known malfunctions at linking goals to urges". People are as bad at looking into hospitals for their own health as for the sake of their parents' health; Kaj didn't actually gain much prestige from feeling guilty about his relaxation time.

We *do* have a status urge. It *does* affect a lot of things. People *do* tend to massively systematically understate it in much the same way that Victorians pretended that sex wasn't everywhere. But that's not the *same* cognitive problem as "Our brain is pretty bad at linking effective behaviors to goals, and will sometimes reward us for just doing things that seem roughly associated with the goal, instead of actions that cause the consequence of the goal being achieved." And our brains not being coherent agents is something that's even more massive than status.

Implication 3: Humans cannot live by urges alone

Like beliefs, goals often get much wackier than urges. I've seen a number of people react to this realization by concluding that they should give up on having goals, and lead an authentic life of pure desire. This wouldn't work any more than giving up on having beliefs. To precisely anticipate how long it takes a ball to fall off a tower, you have to manipulate abstract beliefs about gravitational acceleration. I have an *urge* to drive a car that runs smoothly, but if I didn't also have a *goal* of having a well-maintained car, I would never get around to having it serviced - I have no innate urge to do that.

I really have seen multiple people (some of whom I significantly cared about) malfunctioning as a result of misinterpreting this point. As a stand-alone system for pulling your actions, urges have all kinds of problems. Urges can pull you to stare at an attractive stranger, to walk to the fridge, and even to sprint hard for first base when playing baseball. But unless coupled with goals and far-mode reasoning, urges will not pull you to the component tasks required for any longer-term goods. When I get into my car I have a definite urge for it not to be broken. But absent planning, there would never be a moment when the activity I most desired was to take my car for an oil change. To find and keep a job (let alone a good job), live in a non-pigsty, or learn any skills that are not immediately rewarding, you will probably need goals. *Even though* human goals can easily turn into fashion statements and wishful thinking.

Implication 4: Your agency failures do not imply that your ideals are fake.

Obvious but it needs to be said: People are as bad at looking into hospitals for their own health as for the sake of their parents' health. It doesn't mean that they don't really care about their parents, and it doesn't mean that they don't really care about survival. They would probably run away pretty fast from a tiger, where the goal connected to the urge in an ancestrally more reliable way and hence made them more 'agenty'; and they might fight hard to defend their parents from a tiger too.

There's a very real sense in which our agency failures imply that human beings don't have goals, but this doesn't mean that our ungoaly ideals are any more ungoaly than anything else. Ideals can be more ungoaly because they're sometimes about faraway things or less ancestral things - it's probably *easier* to improve your agency on less idealy goals that link more quickly to urges - but as entities which can look over our own urges and goals and try to improve our agentiness, there's no rule which says

that we can't try to solve some hard problems in this area as well as some easy ones. [2]

Implication 5: You can align urges and goals using the same sort of effort and training that it takes to align anticipations and beliefs.

Although I've heard people saying that we discuss willpower-failure too much on Less Wrong, most of the best stuff I've read has been outside Less Wrong and hasn't made contact with us. For a starting guide to many such skills, see *Eat That Frog* by Brian Tracy [3]. Some basic alignment techniques include:

- Get in the habit of asking "What is the positive consequence?" (Probably more needs to be written about this so that your brain doesn't just answer "I'll be a free software supporter!" which is not what we mean to ask.)
- Andrew Critch's "greedy algorithm": Whenever you catch yourself *really wanting* to do something you *want to want,* immediately reward yourself - by feeding yourself an M&M, or if that's too difficult, immediately pumping your fist and saying "Yes!"
- Whenever you sit down to work, naming a single, high-priority accomplishment for that session. Visualizing that accomplishment, and its positive rewarding consequences, until you *have an urge for it to happen* (instead of just having an urge to log today's hours).

And much the same way that a lot of craziness stems, not so much from "having a wrong model of the world", as "not bothering to have a model of the world", a lot of personal effectiveness isn't so much about "having the right goals" as "bothering to have goals at all" - where unpacking this somewhat Vassarian statement would lead us to ideas like "bothering to have *something* that I check my actions' consequences against, never mind whether or not it's the right thing" or "bothering to have *some* communication-related urge that animates my writing when I write, instead of just sitting down to log a certain number of writing hours during which I feel rewarded from rearranging shiny words".

Conclusion:

Besides an aspiring rationalist, these days I call myself an "aspiring consequentialist".

---

[1] IMO the case of somebody who has the belief "I am good at chess", but instinctively knows to avoid strong chess opponents that would potentially test the belief, ought to be a more central example in our literature than the person who believes they have an dragon in their garage (but instinctively knows that they need to specify that it's invisible, inaudible and generates no carbon dioxide, when we show up with the testing equipment).

[2] See also Ch. 20 of Methods of Rationality:

Professor Quirrell: "Mr. Potter, in the end people all do what they want to do. Sometimes people give names like 'right' to things they want to do, but how could we possibly act on anything but our own desires?"

Harry: "Well, obviously I couldn't act on moral considerations if they lacked the power to move me. But that doesn't mean my wanting to hurt those Slytherins has the power to move me more than moral considerations!"

[3] Thanks to Patri for recommending this book to me in response to an earlier post. It is perhaps not written in the most LW-friendly language -- but, given the value of these skills, I'd recommend wading in and doing your best to pull useful techniques from the somewhat salesy prose. I found much of value there.

The problem with too many rational memes

Like so many of my posts, this one starts with a personal anecdote.

A few weeks ago, my boyfriend was invited to a community event through Meetup.com. The purpose of the meetup was to watch the movie The Elegant Universe and follow up with a discussion. As it turns out, this particular meetup was run by a man who I'll call 'Charlie', the leader of some local Ottawa group designed to help new immigrants to Canada find a social support net. Which, in my mind, is an excellent goal.

Charlie turned out to be a pretty neat guy, too: charismatic, funny, friendly, encouraging everyone to share his or her opinion. Criticizing or shutting out other people's views was explicitly forbidden. It was a diverse group, as he obviously wanted it to be, and by the end everyone seemed to feel pretty comfortable.

My boyfriend, an extremely social being whose main goal in life is networking, was raving by the end about what a neat idea it was to start this kind of group, and how Charlie was a really cool guy. I was the one who should have had fun, since I'm about 100 times more interested in physics than he is, but I was fuming silently.

Why? Because, at various points in the evening, Charlie talked about his own interest in the paranormal and the spiritual, and the books he'd written about it. When we were discussing string theory and its extra dimensions, he made a comment, the gist of which was 'if people's souls go to other dimensions when they die, Grandma could be communicating with you right now from another dimension by tapping spoons.'

Final straw. I bit my tongue and didn't say anything and tried not to show how irritated I was. Which is strange, because I've always been fairly tolerant, fairly agreeable, and very eager to please others. Which is why, when my brain responded 'because he's WRONG and I can't call him out on it because of the no criticism rule!' to the query of 'why are you pissed off?', I was a bit suspicious of that answer.

I do think that Charlie is wrong. I would have thought he was wrong a long time ago. But it wouldn't have bothered me; I know that because I managed to attend various churches for years, even though I thought a lot of their beliefs were wrong, because it didn't matter. They had certain goals in common with me, like wanting to make the world a better place, and there were certain things I could get out of being a community member, like incredibly peaceful experiences of bliss that would reset my always-high stress levels to zero and allow me to survive the rest of the week. Some of the sub-goals they had planned to make the world a better place, like converting people in Third World countries to Christianity, were ones that I thought were sub-optimal or even damaging. But overall, there were more goals we had in common than goals we didn't have in common, and I could, I judged, accomplish those goals we had in common more effectively with them than on my own. And anyway, the church would still be there whether or not I went; if I did go, at least I could talk about stuff like physics with awe and joy (no faking required, thinking about physics does make me feel awe and joy), and increase some of the congregation's scientific literacy a little bit.

Then I stopped going to church, and I started spending more time on Less Wrong, and if I were to try to go back, I'm worried it would be exactly the same as the community meetup. I would sit there fuming because they were wrong and it was socially unacceptable for me to tell them that.

I'm worried because I don't think those feelings are the result of a clearheaded, logical value calculation. Yeah, churches and people who believe in the paranormal waste a lot of money and energy, which could be spent on really useful things otherwise. Yes, that could be a valid reason to reject them, to refuse to be their allies even if some of your goals are the same. But it's not my true rejection. My true rejection is that them being wrong is too annoying for me to want to cooperate. Why? I haven't changed my mind, really, about how much damage versus good I think churches do for the world.

I'm worried that the same process which normalized religion for me is now operating in the opposite direction. I'm worried that a lot of Less Wrong memes, ideas that show membership to the 'rationalist' or 'skeptic' cultures, such as atheism itself, or the idea that religion is bad for humanity...I'm worried that they're sneaking into my head and becoming virulent, that I'm becoming an undiscriminating skeptic. Not because I've been presented with way more evidence for them, and updated on my beliefs (although I have updated on some beliefs based on things I read here), but because that agreeable, eager-to-please subset of my brains sees the Less Wrong community and wants to fit in. There's a part of me that evaluates what I read, or hear people say, or find myself thinking, and imagines Eliezer's response to it. And if that response is negative...ooh, mine had better be negative too.

And that's not strategic, optimal, or rational. In fact, it's preventing me from doing something that might otherwise be a goal for me: joining and volunteering and becoming active in a group that does good things for the Ottawa community. And this transformation has managed to happen without me even noticing, which is a bit scary. I've always thought of myself as someone who was aware of my own thoughts, but apparently not.

Anyone else have the same experience?

The Substitution Principle

Partial re-interpretation of: The Curse of Identity
Also related to: Humans Are Not Automatically Strategic, The Affect Heuristic, The Planning Fallacy, The Availability Heuristic, The Conjunction Fallacy, Urges vs. Goals, Your Inner Google, signaling, etc...

What are the best careers for making a lot of money?

Maybe you've thought about this question a lot, and have researched it enough to have a well-formed opinion. But the chances are that even if you hadn't, some sort of an answer popped into your mind right away. Doctors make a lot of money, maybe, or lawyers, or bankers. Rock stars, perhaps.

You probably realize that this is a difficult question. For one, there's the question of who we're talking about. One person's strengths and weaknesses might make them more suited for a particular career path, while for another person, another career is better. Second, the question is not clearly defined. Is a career with a small chance of making it rich and a large chance of remaining poor a better option than a career with a large chance of becoming wealthy but no chance of becoming rich? Third, whoever is asking this question probably does so because they are thinking about what to do with their lives. So you probably don't want to answer on the basis of what career lets you make a lot of money today, but on the basis of which one will do so in the near future. That requires tricky technological and social forecasting, which is quite difficult. And so on.

Yet, despite all of these uncertainties, some sort of an answer probably came to your mind as soon as you heard the question. And if you hadn't considered the question before, your answer probably didn't take any of the above complications into account. It's as if your brain, while generating an answer, never even considered them.

The thing is, it probably didn't.

Daniel Kahneman, in *Thinking, Fast and Slow*, extensively discusses what I call the Substitution Principle:

> If a satisfactory answer to a hard question is not found quickly, System 1 will find a related question that is easier and will answer it. (Kahneman, p. 97)

System 1, if you recall, is the quick, dirty and parallel part of our brains that renders instant judgements, without thinking about them in too much detail. In this case, the actual question that was asked was ”what are the best careers for making a lot of money”. The question that was actually answered was ”what careers have I come to associate with wealth”.

Here are some other examples of substitution that Kahneman gives:

- *How much would you contribute to save an endangered species?* becomes *How much emotion do I feel when I think of dying dolphins?*
- *How happy are you with your life these days?* becomes *What is my mood right now?*
- *How popular will the president be six months from now?* becomes *How popular is the president right now?*

- *How should financial advisors who prey on the elderly be punished?* becomes *How much anger do I feel when I think of financial predators?*

All things considered, this heuristic probably works pretty well most of the time. The easier questions are not meaningless: while not completely accurate, their answers are still generally correlated with the correct answer. And a lot of the time, that's good enough.

But I think that the Substitution Principle is also the mechanism by which most of our biases work. In The Curse of Identity, I wrote:

> In each case, I thought I was working for a particular goal (become capable of doing useful Singularity work, advance the cause of a political party, do useful Singularity work). But as soon as I set that goal, my brain automatically and invisibly re-interpreted it as the goal of doing something that gave the impression of doing prestigious work for a cause (spending all my waking time working, being the spokesman of a political party, writing papers or doing something else few others could do).

As Anna correctly pointed out, I resorted to a signaling explanation here, but a signaling explanation may not be necessary. Let me reword that previous generalization: As soon as I set a goal, my brain asked itself how that goal might be achieved, realized that this was a difficult question, and substituted it with an easier one. So ”how could I advance X” became ”what are the kinds of behaviors that are commonly associated with advancing X”. That my brain happened to pick the most prestigious ways of advancing X might be simply because prestige is often correlated with achieving a lot.

Does this exclude the signaling explanation? Of course not. My behavior is probably still driven by signaling and status concerns. One of the mechanisms by which this works might be that such considerations get disproportionately taken into account when choosing a heuristic question. And a lot of the examples I gave in The Curse of Identity seem hard to justify without a signaling explanation. But signaling need not to be the *sole* explanation. Our brains may just resort to poor heuristics a lot.

Some other biases and how the Substitution Principle is related to them (many of these are again borrowed from Thinking, Fast and Slow):

The Planning Fallacy: ”How much time will this take” becomes something like ”How much time did it take for me to get this far, and many times should that be multiplied to get to completion.” (Doesn't take into account unexpected delays and interruptions, waning interest, etc.)

The Availability Heuristic: ”How common is this thing” or ”how frequently does this happen” becomes ”how easily do instances of this come to mind”.

Over-estimating your own share of household chores: ”What fraction of chores have I done” becomes ”how many chores do I remember doing, as compared to the amount of chores I remember my partner doing.” (You will naturally remember more of the things that you've done than that somebody else has done, possibly when you weren't even around.)

Being in an emotionally ”cool” state and over-estimating your degree of control in an emotionally ”hot” state (angry, hungry, sexually aroused, etc.):

”How well could I resist doing X in that state” becomes ”how easy does resisting X feel like now”.

The Conjunction Fallacy: ”What's the probability that Linda is a feminist” becomes ”how representative is Linda of my conception of feminists”.

People voting for politicians for seemingly irrelevant reasons: ”How well would this person do his job as a politician” becomes ”how much do I like this person.” (A better heuristic than you might think, considering that we like people who like us, owe us favors, resemble us, etc. - in the ancestral environment, supporting the leader you liked the most was probably a pretty good proxy for supporting the leader who was most likely to aid you in return.)

And so on.

The important point is to **learn to recognize the situations where you're confronting a difficult problem, and your mind gives you an answer right away.** If you don't have extensive expertise with the problem – or even if you do – it's likely that the answer you got wasn't actually the answer to the question you asked. So before you act, stop to consider what heuristic question your brain might actually have used, and whether it makes sense given the situation that you're thinking about.

This involves three skills: first **recognizing a problem as a difficult one**, then **figuring out what heuristic you might have used**, and finally **coming up with a better solution**. I intend to develop something on how to taskify those skills, but if you have any ideas for how that might be achieved, let's hear them.

The Singularity Institute's Arrogance Problem

I intended Leveling Up in Rationality to communicate this:

> Despite worries that extreme rationality isn't that great, I think there's reason to hope that it *can* be great if some other causal factors are flipped the right way (e.g. mastery over akrasia). Here are some detailed examples I can share because they're from my own life...

But some people seem to have read it and heard this instead:

> I'm super-awesome. Don't you wish you were more like me? Yay rationality!

This failure (on my part) fits into a larger pattern of the Singularity Institute *seeming* too arrogant and (perhaps) *being* too arrogant. As one friend recently told me:

> At least among Caltech undergrads and academic mathematicians, it's taboo to toot your own horn. In these worlds, one's achievements speak for themselves, so whether one is a Fields Medalist or a failure, one gains status purely passively, and must appear not to care about being smart or accomplished. I think because you and Eliezer don't have formal technical training, you don't instinctively grasp this taboo. Thus Eliezer's claim of world-class mathematical ability, in combination with his lack of technical publications, make it hard for a mathematician to take him seriously, because his social stance doesn't pattern-match to anything good. Eliezer's arrogance as evidence of technical cluelessness, was one of the reasons I didn't donate until I met [someone at SI in person]. So for instance, your boast that at SI discussions "everyone at the table knows and applies an insane amount of all the major sciences" would make any Caltech undergrad roll their eyes; your standard of an "insane amount" seems to be relative to the general population, not relative to actual scientists. And posting a list of powers you've acquired doesn't make anyone any more impressed than they already were, and isn't a high-status move.

So, I have a few questions:

1. What are the most egregious examples of SI's arrogance?
2. On which subjects and in which ways is SI too arrogant? Are there subjects and ways in which SI isn't arrogant enough?
3. What should SI do about this?

Completeness, incompleteness, and what it all means: first versus second order logic

First order arithmetic is incomplete. Except that it's also complete. Second order arithmetic is more expressive - except when it's not - and is also incomplete and also complete, except when it means something different. Oh, and full second order-logic might not really be a logic at all. But then, first order logic has no idea what the reals and natural numbers are, especially when it tries to talk about them.

That was about the state of my confusion, and I set out to try and clear it up. Here I'll try and share an understanding of what is really going on with first and second order logic and why they differ so radically. It will be deliberately informal, so I won't be distinguishing between functions, predicates and subsets, and will be using little notation. It'll be exactly what I wish someone had told me before I started looking into the whole field.

Meaningful Models

An old man starts talking to you about addition, subtraction and multiplication, and how they interact. You assume he was talking about the integers; turns out he means the rational numbers. The integers and the rationals are both models of addition, subtraction and multiplication, in that they obey all the properties that the old man set out. But notice though he had the rationals in mind, he didn't mention them at all, he just listed the properties, and the rational numbers turned out, very non-coincidentally, to obey them.

These models are generally taken to give meaning to the abstract symbols in the axioms - to give semantics to the syntax. In this view, "for all x,y xy=yx" is a series of elegant squiggles, but once we have the model of the integers (or the rationals) in mind, we realise that this means that multiplication is commutative.

Similarly, models can define the "truth" of sentences. Consider the following sentences:

(1) 2 has a multiplicative inverse.
(2) There exists a number that squares to -1.
(3) 2 is not equal to zero.
(4) If $a+b=0$, then $a^2=b^2$.
(5) No number is equal to zero.

Are these true? You and the old man would disagree on (1), with him saying yes and you saying no - your models have enabled you to attach truth-values to the statement. You would both claim (2) is false, but there are other models - such as the complex numbers - where it is true. You would both claim (3) is true, but there are other models - such as the field with two elements - where it is false. So truth is model dependent.

The last two are interesting, because it turns out that (4) is true in every model, and (5) is false. Statements like (4) and "not (5)" that are true in every model are called valid. Since they are independent of the choice of models, these statements are, in a certain sense, true from pure syntax. Both these statement can also be deduced purely from the axioms. It would be nice if all valid statements could also be deduced. But only first order logic allows this.

What would also be nice if you could agree on the model you're using. Maybe the old man could add "2 (and every non-zero number) has a multiplicative inverse", giving the field axioms, and ruling your integers right out. But there are still many fields - the rationals, the reals, the complex numbers, and many in between. But maybe with a few more axioms, you could really narrow thing downs, and treat the axioms and the model interchangeably. But only second order logic allows *this*.

First order fun

First order theories are those where you can quantify over the basic objects in your theory, and phrase statements like "all Greeks enjoy dancing" and "there exists a blind millionaire". This distinguishes them from second order theories where you can quantify over higher order objects (predicates and functions), phrasing sentiments like "all nationalities are equal" or "there exists a dominant social class" - in first order logic with humans as basic objects, nationality and social class are predicates, and you can't quantify over them. You don't appreciate how limiting first order logic can be until you've worked with it a while; nevertheless, it's a good logic to start with and possesses certain key properties not shared by higher order logics. Let's start with the most famous result, the incompleteness theorem.

Gödel's incompleteness theorem

Gödel's (first) incompleteness theorem, is a theorem about an arithmetic but also (implicitly) about a model. The implicit model is the natural numbers: any arithmetic that can model them suffers from the incompleteness theorem. But it is not really about any model, beyond that requirement: it's an intrinsic limitation of the system.

Let's assume we have the usual (first order) Peano axioms for ordering, addition and multiplication. We also need an infinity of axioms to define induction, but that isn't as bad as it seem: given a specific sentence, it's easy to check whether or not it's an axiom, in a fast and efficient way. To nobody's surprise, the natural number are a model of first order Peano arithmetic.

And inside this model, we can construct the Gödel sentence G, which is equivalent with "there is no proof of G". By 'proof', we mean a natural number n that is the Gödel number that encodes a proof of G. Obviously, Peano arithmetic cannot prove G without being inconsistent; but this is precisely what G is saying, so we can actually see that G is true. And hence that "not G" cannot be provable if our arithmetic is consistent. This is the incompleteness theorem: neither G nor "not G" are provable, so the proof system is "incomplete".

Gödel's completeness theorem

Enough with *in*completeness; what about Gödel completeness theorem? Unlike the previous theorem, this is a statement about about the axiomatic system and *all* of its models. It simply says that if a sentence is valid (true in every model) for a first order theory, then it can be proved from the axioms. This provides a bridge between the semantic concept of "true" (true in every model) and the syntactic concept of provable (can be proved by these formal manipulations). It also implies that we can enumerate all the sentences that are valid in a first order system, simply by enumerating all the proofs.

Where does this leave the Gödel sentence G? We've seen we can't prove it from the axioms, hence it cannot be true in all models. Therefore there must exist a model N' of first order Peano arithmetic in which G is false. What does that mean? G claims that "there does not exist a number n with (certain properties)", so if G is false, such an n does exist. Now we know (because we've constructed it that way) that if that n were a natural number, then those (certain properties) means that it must encode a proof of G. Since there is no proof of G, n cannot be a natural number, but must be an extra, a non-standard number, from beyond our usual universe. This also means that those (certain properties) do not capture what we thought they did: they only mean "encodes a proof of G" for the standard natural numbers.

This seems somewhat troubling, that Peano arithmetic would admit two distinct models and fail to say what we thought it said; but it gets worse.

The Löwenheim–Skolem theorem

The Löwenheim–Skolem theorem says that if any countable first order theory (such as Peano arithmetic) has an infinite model, then it has a model for every size ("cardinality") of infinity. Therefore first order Peano arithmetic has to have many, many models; at a minimum, it has to have a model of same size as the reals.

This also means that no matter how much first order information we add to Peano arithmetic, we cannot restrict it to only being about the natural numbers; the models and the axioms can never be interchangeable. But it gets still worse.

Uncountable models of Peano arithmetic are bad enough, but it turns out that Peano arithmetic also has many other countable models - models of same size as the natural numbers, but weirdly different. Weirdly is an apt summation of these models where neither multiplication nor addition can be computed.

So first order Peano arithmetic is not really about the natural numbers at all - but about the natural numbers and all these strange countable and uncountable models that we can't really describe. Maybe second order theories will do better?

Second order scariness

In second order theories, we can finally do what we've been itching to do: apply the existential and universal quantifiers to predicates, functions, sets of numbers and objects of that ilk. We can triumphantly toss away the infinitely many axioms needed to define induction in first order Peano arithmetic, and replace them with a simple:

- "Every (non-empty) set of numbers has a least element."

This, as every schoolboy should know, is enough to uniquely define the natural numbers. Or is it?

The importance of semantics

That sentence remains a series of squiggles until we've decided what those squiggles actually *mean*. This takes on an extra importance in second order logic that it didn't have in first order. When we say "every set of numbers", what do we mean? In terms of meaning and models, what models are we going to be considering?

The first idea is the obvious one: "every set of numbers" means, duh, "*every* set of numbers". When we specify a model, we'll give the 'universe of discourse', the basic objects (maybe the natural numbers or the reals), and the quantifier 'every' will range over all possible subsets of this universe (every subset of the natural or real numbers). This is called the full or standard semantics for second order arithmetic, and the models are called full models.

But note what we have done here: we have brought in extra information to clarify what we are talking about. In order to defined full semantics, we've had to bring set theory into the mix, to define all these subset. This caused Quine to accuse second order logic of not being a logic at all, but a branch of set theory.

Also with all these Russell Paradoxes flying around, we might be a bit wary of jumping immediately into the 'every set' semantics. Maybe instead of defining a model by just giving the 'universe of discourse', we would want to also define the subsets we are interested in, listing explicitly all the sets we are going to allow the quantifiers to range over.

But this could get ridiculous - do we really want a model which includes the natural numbers, but only allows us to quantify over the sets {1, 7, 13908}, {0} and the empty set? We can define, for instance, what an even number is (a number that is equal to two times something), and so why can't we get the set of even numbers into our model?

Henkin semantics

We really want our 'universe of quantifiable sets' to include any set we can define. It turns out this something we can get from inside second-order logic, by using the comprehension axioms. They roughly say that "any set/predicate we can define, is in the universe of sets/predicates we can quantify over". There are infinitely many such comprehension axioms, covering each definition.

Then Henkin semantics is second-order logic, with all the comprehension axioms, and no further restrictions on the possible models. These 'Henkin models' will have both a defined universe of discourse (the list of basic objects in the model we can quantify over) and a defined 'universe of sets/predicates' (a list of sets of basic objects that we can quantify over), with the comprehension axioms making sure they are compatible. Though they are called 'Henkin semantics', they could really be called 'Henkin syntaxes', since we aren't giving any extra restrictions on the models apart from the internal axioms.

It should be noted that a full model (where the 'universe of sets' include all possible subsets) automatically obeys the comprehension axioms, since it can quantify over every set. So every full model in a Henkin model, and it might seem that Henkin semantics are a simple extension of full semantics, and that they have a greater 'expressive power'. Few things could be further from the truth.

First or second order?

If the old man of previously had claimed "every number is even", and, when challenged with "3" had responded "3 is not a number", you might be justified in questioning his grasp of the meaning of 'every' and 'number'. Similarly, if he had said "every (non-empty) set of *integers* has a least element," and when challenged with "the negative integers" had responded "that collection is not a set", you would also question his use of 'every'.

Similarly, since Henkin semantics allows us to restrict the meaning of "every set", depending on the model under consideration, statements such as "every set is blah" are much weaker in Henkin semantics than in full semantics. For instance, take the axioms for an ordered field, and add:

- "Every (non-empty) bounded set has a supremum"

As every schoolgirl should know, this is enough to model the real numbers... in full semantics. But in Henkin semantics, 'every bounded set' can mean 'every definable bounded set' and we can take the 'definable reals' as a model: the supremum of a definable set is definable. And this does not include all the real numbers; for a start, the definable reals are countable, so there are far fewer of them than there are reals.

This may seem a feature, rather than a bug. What are these strange, 'non-definable reals' that clutter up the place; why would we want them anyway? But the definable reals are just one Henkin model of these axioms, there are others perfectly valid models, including the reals themselves. So these axioms have not allowed us, in Henkin semantics, to pick out one particular model.

This seems familiar: the first order Peano axioms also failed to specify the natural numbers. The familiarity is not an illusion: Henkin semantics is actually a first order theory (a 'many sorted' one, where some classes of objects have different properties). Hence the completeness theorem still applies: any result true in every Henkin model, can be proved from the basic axioms. But this is not much help if we have many models, and unfortunately the Löwenheim–Skolem theorem also still applies: if we have one infinite model, we have many, many others. So not only do we have the countable 'definable reals' and the reals themselves as models but larger 'hyperreals' and 'superreals' with many more elements to them.

Skolem's paradox

In fact, Henkin semantics can behave much worse than standard first order logic, as it can express more. Express more - but ultimately, not mean more. For instance, in second order language, we can express the sentence "there exists an uncountable set". We could start by defining an infinite set as one with a one-to-one correspondence with a strict subset of itself, *á la* Cantor. We could define an uncountable infinite set as one that has a subset that is also infinite, but that doesn't

have a one-to-one correspondence with it (the subset is of lower cardinality). There are other, probably better, ways of phrasing the same concept, but that will do for here.

Then basic second order logic with Henkin semantics and the additional axiom "there exists an uncountable set" certainly has a model: the reals, for instance. Then by the Löwenheim–Skolem theorem, it must have a countable model.

Wait a moment there. A logic that asserts the existence of an uncountable set... has a countable model? This was Skolem's paradox, and one of his arguments against first order logic. The explanation for the paradox involves those one-to-one correspondences mentioned above. An uncountable set is an infinite set without any one-to-one functions to any countable set. But in a Henkin model 'any one-to-one function' means 'any one to one function on the list of allowable functions in this model'. So the 'uncountable set' in the countable model is, in fact, countable: it has one-to-one functions to other countable set. But all these functions are banned from the Henkin model, so the model cannot see, internally, that that set is actually countable.

So we can *express* a lot of statements in Henkin semantics - "every bounded set has a supremum", "there exists an uncountable set" - but these don't actually *mean* what we thought they did.

Full second order semantics

Having accepted the accusations of sneaking in set theory, and the disturbing fact that we had to bring in meaning and semantics (by excluding a lot of potential models), rather than relying on the syntax... what can we do with full second order semantics?

Well, for start, finally nail down the natural numbers and the reals. With second order Peano arithmetic, including the second order induction axiom "every (non-empty) set of numbers has a least element", we know that we have only one (full) model: the natural numbers. Similarly, if we have the axioms for an ordered field and toss in "every bounded set has a supremum", then the reals are the only full model that stands up.

This immediately implies that full second order semantics are not complete, unlike Hekin semantics and first order theories. We can see this from the incompleteness result (though don't confuse incompleteness with non-completeness). Take second order Peano arithmetic. This has a Gödel statement G which is true but unprovable. But there is only one model of second order Peano arithmetic! So G is both unprovable and true in every model for the theory.

It may seem surprising that completeness fails for full semantics: after all, it is true in Henkin semantics, and every full model is also a Henkin model, so how can this happen? It derives from the restriction of possible models: completeness means that every sentence that is true in every Henkin model, must be provable. That does not mean that every sentence that is true in every full model, must also be provable. The G sentence is indeed false in some models - but only in Henkin models that are not full models.

The lack of completeness means that the truths of second order logic cannot be enumerated - it has no complete proof procedure. This causes some to reject full second order logic on these grounds. Others argued that completeness is not the important factor, but rather decidability: listing all the provable statements might be light entertainment, but what we really want is an algorithm to be able to prove (or disprove) any given sentence. But the Church-Turing theorem demonstrates that this cannot be done, in either first or second order logic: hence neither system can claim to be superior in this respect.

Higher-order logic within full second order logic

Higher order logic is the next step up - quantifying over predicates of predicates, functions of functions. This would seem to make everything more complicated. However there is a result due to Hintikka that any sentence in full higher order logic can be shown to be equivalent (in an effective manner) with a sentence in full second order logic, using many-sorting. So there is, in a certain sense, no need to go beyond, and the important debate is between first order and full second order logic.

Conclusion

So, which logic is superior? It depends to some extent on what we need it for. Anything provable in first order logic can be proved in second order logic, so if we have a choice of proofs, picking the first order one is the better option. First order logic has more pleasing internal properties, such as the completeness theorem, and one can preserve this in second order via Henkin semantics without losing the ability to formally express certain properties. Finally, one needs to make use of set theory and semantics to define full second order logic, while first order logic (and Henkin semantics) get away with pure syntax.

On the other hand, first order logic is completely incapable of controlling its infinite models, as they multiply, uncountable and generally incomprehensible. If rather that looking at the logic internally, we have a particular model in mind, we have to use second order logic for that. If we'd prefer not to use infinitely many axioms to express a simple idea, second-order logic is for us. And if we really want to properly express ideas like "every (non-empty) set has a least element", "every analytic function is uniquely defined by its power series" - and not just express them, but have them mean what we want them to mean - then full second order logic is essential.

***EDIT**: an addendum addresses the problem of using set theory (a first order theory) to define second order logic.*

So You Want to Save the World

This post is very out-of-date. See MIRI's research page for the current research agenda.

So you want to save the world. As it turns out, the world cannot be saved by caped crusaders with great strength and the power of flight. No, the world must be saved by mathematicians, computer scientists, and philosophers.

This is because the creation of machine superintelligence this century will determine the future of our planet, and in order for this "technological Singularity" to go *well* for us, we need to solve a particular set of technical problems in mathematics, computer science, and philosophy *before* the Singularity happens.

The best way for most people to save the world is to donate to an organization working to solve these problems, an organization like the Singularity Institute or the Future of Humanity Institute.

Don't underestimate the importance of donation. You can do more good as a philanthropic banker than as a charity worker or researcher.

But if you *are* a capable researcher, then you may also be able to contribute by working directly on one or more of the open problems humanity needs to solve. If so, read on...

Preliminaries

At this point, I'll need to assume some familiarity with the subject matter. If you haven't already, take a few hours to **read these five articles**, and then come back:

1. Yudkowsky (2008a)
2. Sandberg & Bostrom (2008)
3. Chalmers (2010)
4. Omohundro (2011)
5. Armstrong et al. (2011)

Or at the very least, read my shorter and more accessible summary of the main points in my online book-in-progress, *Facing the Singularity*.

Daniel Dewey's highly compressed summary of several key points is:

> Hardware and software are improving, there are no signs that we will stop this, and human biology and biases indicate that we are far below the upper limit on intelligence. Economic arguments indicate that most AIs would act to become more intelligent. Therefore, intelligence explosion is very likely. The apparent diversity and irreducibility of information about "what is good" suggests that value is complex and fragile; therefore, an AI is unlikely to have any significant overlap with human values if that is not engineered in at significant cost. Therefore, a bad AI explosion is our default future.

> The VNM utility theorem suggests that there is some formally stated goal that we most prefer. The CEV thought experiment suggests that we could program a metaethics that would generate a good goal. The Gandhi's pill argument indicates that goal-preserving self-improvement is possible, and the reliability of formal proof suggests that long chains of self-improvement are possible. Therefore, a good AI explosion is likely possible.

Next, I need to make a few important points:

1. Defining each problem is part of the problem. As Bellman (1961) said, "the very construction of a precise mathematical statement of a verbal problem is itself a problem of major difficulty." Many of the problems related to navigating the Singularity have not yet been stated with mathematical precision, and the need for a precise statement of the problem is *part* of the problem. But there is reason for optimism. Many times, particular heroes have managed to formalize a previously fuzzy and mysterious concept: see Kolmogorov on complexity and simplicity (Kolmogorov 1965; Grunwald & Vitanyi 2003; Li & Vitányi 2008), Solomonoff on induction (Solomonoff 1964a, 1964b; Rathmanner & Hutter 2011), Von Neumann and Morgenstern on rationality (Von Neumann & Morgenstern 1947; Anand 1995), and Shannon on information (Shannon 1948; Arndt 2004).

2. The nature of the problem space is unclear. Which problems will biological humans need to solve, and which problems can a successful Friendly AI (FAI) solve on its own (perhaps with the help of human uploads it creates to solve the remaining open problems)? Are Friendly AI (Yudkowsky 2001) and CEV (Yudkowsky 2004) coherent ideas, given the confused nature of human "values"? Should we aim instead for something like Oracle AI (Armstrong et al. 2011)? Which problems are we unable to state with precision because they are irreparably confused, and which problems are we unable to state due to a lack of insight?

3. Our intervention priorities are unclear. There are a limited number of capable researchers who will work on these problems. Which are the most important problems they should be working on, if they are capable of doing so? Should we focus on "control problem" theory (FAI, AI-boxing, oracle AI, etc.), or on strategic considerations (differential technological development, methods for raising the sanity waterline, methods for bringing more funding to existential risk reduction and growing the community of x-risk reducers, reducing the odds of AI arms races, etc.)? Is AI more urgent than other existential risks, especially synthetic biology? Is research the most urgent thing to be done, or should we focus on growing the community of x-risk reducers, raising the sanity waterline, bringing in more funding for x-risk reduction, etc.? Can we make better research progress in the next 10 years if we work to improve sanity and funding for 7 years and *then* have the resources to grab more and better researchers, or can we make better research progress by focusing on research now?

Problem Categories

There are many ways to categorize our open problems; I'll divide them into three groups:

Safe AI Architectures. This may include architectures for securely confined or "boxed" AIs (Lampson 1973), including Oracle AIs, and also AI architectures capable of using a safe set of goals (resulting in Friendly AI).

Safe AI Goals. What could it mean to have a Friendly AI with "good" goals?

Strategy. How do we predict the future and make recommendations for differential technological development? Do we aim for Friendly AI or Oracle AI or both? Should we focus on growing support now, or do we focus on research? How should we interact with the public and with governments?

The list of open problems on this page is *very* preliminary. I'm sure there are many problems I've forgotten, and many problems I'm unaware of. Probably *all* of the problems are stated poorly: this is only a "first step" document. Certainly, all listed problems are described at a very "high" level, far away (so far) from mathematical precision, and can themselves be broken down into several and often *dozens* of subproblems.

Safe AI Architectures

Is "rationally-shaped" transparent AI the only potentially safe AI architecture? Omohundro (2007, 2008, 2011) describes "rationally shaped" AI as AI that is as economically rational as possible given its limitations. A rationally shaped AI has beliefs and desires, its desires are defined by a utility function, and it seeks to maximize its expected utility. If an AI doesn't use a utility function, then it's hard to predict its actions, including whether they will be "friendly." The same problem can arise if the decision mechanism or the utility function is not transparent to humans. At least, this *seems* to be the case, but perhaps there are strong attractors that would allow us to predict friendliness even without the AI having a transparent utility function, or even a utility function at all? Or, perhaps a new decision theory could show the way to a different AI architecture that would allow us to predict the AI's behavior without it having a transparent utility function?

How can we develop a reflective decision theory? When an agent considers radical modification of its own decision mechanism, how can it ensure that doing so will keep constant or increase its expected utility? Yudkowsky (2011a) argues that current decision theories stumble over Löb's Theorem at this point, and that a new, "reflectively consistent" decision theory is needed.

How can we develop a timeless decision theory with the bugs worked out? Paradoxes like Newcomb's Problem (Ledwig 2000) and Solomon's Problem (Gibbard & Harper 1978) seem to show that neither causal decision theory nor evidential decision theory is ideal. Yudkowsky (2010) proposes an apparently superior alternative, timeless decision theory. But it, too, has bugs that need to be worked out, for example the "5-and-10 problem" (described here by Gary Drescher, who doesn't use the 5-and-10 example illustration).

How can we modify a transparent AI architecture to have a utility function over the external world? Reinforcement learning can only be used to define agents whose goal is to maximize expected rewards. But this doesn't match human goals, so advanced reinforcement learning agents will diverge from our wishes. Thus, we need a

class of agents called "value learners" (Dewey 2011) that "can be designed to learn and maximize any initially unknown utility function" (see Hibbard 2011 for clarifications). Dewey's paper, however, is only the first step in this direction.

How can an agent keep a stable utility function through ontological shifts? An agent's utility function may refer to states of, or entities within, its ontology. As De Blanc (2011) notes, "If the agent may upgrade or replace its ontology, it faces a crisis: the agent's original [utility function] may not be well-defined with respect to its new ontology." De Blanc points toward some possible solutions for these problems, but they need to be developed further.

How can an agent choose an ideal prior? We want a Friendly AI's model of the world to be as accurate as possible so that it successfully does friendly things if we can figure out how to give it friendly goals. Solomonoff induction (Li & Vitanyi 2008) may be our best formalization of induction yet, but it could be improved upon.

First, we may need to solve the problem of observation selection effects or "anthropic bias" (Bostrom 2002b): even an agent using a powerful approximation of Solomonoff induction may, due to anthropic bias, make radically incorrect inferences when it does not encounter sufficient evidence to update far enough away from its priors. Several solutions have been proposed (Neal 2006; Grace 2010;, Armstrong 2011), but none are as yet widely persuasive.

Second, we need improvements to Solomonoff induction. Hutter (2009) discusses many of these problems. We may also need a version of Solmonoff induction in second-order logic because second-order logic with binary predicates can simulate higher-order logics with nth-order predicates. This kind of Solomonoff induction would be able to imagine even, for example, hypercomputers and time machines.

Third, we would need computable approximations for this improvement to Solomonoff induction.

What is the ideal theory of how to handle logical uncertainty? Even an AI will be uncertain about the true value of certain logical propositions or long chains of logical reasoning. What is the best way to handle this problem? Partial solutions are offered by Gaifman (2004), Williamson (2001), and Haenni (2005), among others.

What is the ideal computable approximation of perfect Bayesianism? As explained elsewhere, we want a Friendly AI's model of the world to be as accurate as possible. Thus, we need ideal computable theories of priors and of logical uncertainty, but we also need computable approximations of Bayesian inference. Cooper (1990) showed that inference in unconstrained Bayesian networks is NP-hard, and Dagum & Luby (1993) showed that the corresponding approximation problem is also NP-hard. The most common solution is to use randomized sampling methods, also known as "Monte Carlo" algorithms (Robert & Casella 2010). Another approach is variational approximation (Wainwright & Jordan 2008), which works with a simpler but similar version of the original problem. Another approach is called "belief propagation" — for example, loopy belief propagation (Weiss 2000).

Can we develop a safely confined AI? Can we develop Oracle AI? One approach to constraining a powerful AI is to give it "good" goals. Another is to externally constrain it, creating a "boxed" AI and thereby "leakproofing the singularity" (Chalmers 2010). A *fully* leakproof singularity is impossible or pointless: "For an AI system to be useful... to us at all, it must have some effects on us. At a minimum, we must be able to observe it." Still, there may be a way to constrain a superhuman AI

such that it is useful but not dangerous. Armstrong et al. (2011) offer a detailed proposal for constraining an AI, but there remain many worries about how safe and sustainable such a solution is. The question remains: Can a superhuman AI be safely confined, and can humans managed to safely confine *all* superhuman AIs that are created?

What convergent AI architectures and convergent instrumental goals can we expect from superintelligent machines? Omohundro (2008, 2011) argues that we can expect that "as computational resources increase, there is a natural progress through stimulus-response systems, learning systems, reasoning systems, self-improving systems, to fully rational systems," and that for rational systems there are several convergent instrumental goals: self-protection, resource acquisition, replication, goal preservation, efficiency, and self-improvement. Are these claims true? Are there additional convergent AI architectures or instrumental goals that we can use to predict the implications of machine superintelligence?

Safe AI Goals

Can "safe" AI goals only be derived from contingent "desires" and "goals"? Might a single procedure for responding to goals be uniquely determined by reason? A natural approach to selecting goals for a Friendly AI is to ground them in an extrapolation of current human goals, for this approach works even if we assume the naturalist's standard Humean division between motives and reason. But might a sophisticated Kantian approach work, such that some combination of decision theory, game theory, and algorithmic information theory provides a uniquely dictated response to goals? Drescher (2006) attempts something like this, though his particular approach seems to fail.

How do we construe a utility function from what humans "want"? A natural approach to Friendly AI is to program a powerful AI with a utility function that accurately represents an extrapolation of what humans want. Unfortunately, humans do not seem to have coherent utility functions, as demonstrated by the neurobiological mechanisms of choice (Dayan 2011) and behavioral violations of the axioms of utility theory (Kahneman & Tversky 1979). Economists and computer scientists have tried to extract utility theories from human behavior with choice modelling (Hess & Daly 2010) and preference elicitation (Domshlak et al. 2011), but these attempts have focused on extracting utility functions over a narrow range of human preferences, for example those relevant to developing a particular decision support system. We need new more powerful and universal methods for preference extraction. Or, perhaps we must allow actual humans to reason about their own preferences for a very long time until they reach a kind of "reflective equilibrium" in their preferences (Yudkowsky 2004). The best path may be to upload a certain set of humans, which would allow them to reason through their preferences with greater speed and introspective access. Unfortunately, the development of human uploads may spin off dangerous neuromorphic AI before this can be done.

How should human values be extrapolated? Value extrapolation is an old subject in philosophy (Muehlhauser & Helm 2011), but the major results of the field so far have been to show that certain approaches *won't* work (Sobel 1994); we still have no value extrapolation algorithms that might plausibly work.

Why extrapolate the values of humans alone? What counts as a human? Do values converge if extrapolated? Would the choice to extrapolate the values of humans alone be an unjustified act of speciesism, or is it justified because humans are special in some way — perhaps because humans are the only beings who can reason about their own preferences? And what counts as a human? The problem is more complicated than one might imagine (Bostrom 2006; Bostrom & Sandberg 2011). Moreover, do we need to scan the values of all humans, or only some? These problems are less important if values converge upon extrapolation for a wide variety of agents, but it is far from clear that this is the case (Sobel 1999, Doring & Steinhoff 2009).

How do aggregate or assess value in an infinite universe? What can we make of other possible laws of physics? Our best model of the physical universe predicts that the universe is spatially infinite, meaning that all possible "bubble universes" are realized an infinite number of times. Given this, how do we make value calculations? The problem is discussed by Knobe (2006) and Bostrom (2009), but more work remains to be done. These difficulties may be exacerbated if the universe is infinite in a stronger sense, for example if all possible mathematical objects exist (Tegmark 2005).

How should we deal with normative uncertainty? We may not solve the problems of value or morality in time to build Friendly AI. Perhaps instead we need a theory of how to handle this normative uncertainty. Sepielli (2009) and Bostrom (2009) have made the initial steps, here.

Is it possible to program an AI to do what is "morally right" rather than give it an extrapolation of human goals? Perhaps the only way to solve the Friendly AI problem is to get an AI to do moral philosophy and come to the correct answer. But perhaps this exercise would only result in the conclusion that our moral concepts are incoherent (Beavers 2011).

Strategy

What methods can we use to predict technological development? Predicting progress in powerful technologies (AI, synthetic biology, nanotechnology) can help us decide which existential threats are most urgent, and can inform our efforts in differential technological development (Bostrom 2002a). The stability of Moore's law may give us limited predictive hope (Lundstrom 2003; Mack 2011), but in general we have no proven method for long-term technological forecasting, including expert elicitation (Armstrong 1985; Woudenberg 1991; Rowe & Wright 2001) and prediction markets (Williams 2011). Nagy's performance curves database (Nagy 2010) may aid our forecasting efforts, as may "big data" in general (Weinberger 2011).

Which kinds of differential technological development should we encourage, and how? Bostrom (2002) proposes a course of *differential technological development*: "trying to retard the implementation of dangerous technologies and accelerate implementation of beneficial technologies, especially those that ameliorate the hazards posed by other technologies." Many examples are obvious: we should retard the development of technologies that pose an existential risk, and accelerate the development of technologies that help protect us from existential risk, such as vaccines and protective structures. Some potential applications are less obvious.

Should we accelerate the development of whole brain emulation technology so that uploaded humans can solve the problems of Friendly AI, or will the development of WBEs spin off dangerous neuromorphic AI first? (Shulman & Salamon 2011)

Which open problems are safe to discuss, and which are potentially highly dangerous. There was a recent debate on whether a certain scientist should publish his discovery of a virus that "could kill half of humanity." (The answer in this case was "no.") The question of whether to publish results is particularly thorny when it comes to AI research, because most of the work in the "Safe AI Architectures" section above would, if completed, bring us closer to developing *both* uFAI and FAI, but in particular it would make it easier to develop uFAI. Unfortunately, it looks like that work must be done to develop *any* kind of FAI, while if it is *not* done then *only* uFAI can be developed (Dewey 2011).

What can we do to reduce the risk of an AI arms race? AGI is, in one sense, a powerful weapon for dominating the globe. Once it is seen by governments as a feasible technology goal, we may predict an arms race for AGI. Shulman (2009) gives several reasons to recommend "cooperative control of the development of software entities" over other methods for arms race risk mitigation, but these scenarios require more extensive analysis.

What can we do to raise the "sanity waterline," and how much will this help? The Singularity Institute is a strong advocate of rationality training, in part so that both AI safety researchers and supporters of x-risk reduction can avoid the usual thinking failures that occur when thinking about those issues (Yudkowsky 2008b). This raises the question of how well rationality can be taught, and how much difference it will make for existential risk reduction.

What can we do to attract more funding, support, and research to x-risk reduction and to specific sub-problems of successful Singularity navigation? Much is known about how to raise funding (Oppenheimer & Olivola 2010) and awareness (Kotler & Armstrong 2009), but applying these principles is always a challenge, and x-risk reduction may pose unique problems for these tasks.

Which interventions should we prioritize? There are limited resources available for existential risk reduction work, and for AI safety research in particular. How should these resources be allocated? Should the focus be on direct research, or on making it easier for a wider pool of researchers to contribute, or on fundraising and awareness-raising, or on other types of interventions?

How should x-risk reducers and AI safety researchers interact with governments and corporations? Governments and corporations are potential sources of funding for x-risk reduction work, but they may also endanger the x-risk reduction community. AI development labs will be unfriendly to certain kinds of differential technological development advocated by the AI safety community, and governments may face pressures to nationalize advanced AI research groups (including AI safety researchers) once AGI draws nearer.

How can optimal philanthropists get the most x-risk reduction for their philanthropic buck? Optimal philanthropists aim not just to make a difference, but to make the *most possible* positive difference. Bostrom (2011) makes a good case for existential risk reduction as optimal philanthropy, but more detailed questions remain. *Which* x-risk reduction interventions and organizations should be funded? Should new

organizations be formed, or should resources be pooled in one or more of the existing organizations working on x-risk reduction?

How does AI risk compare to other existential risks? Yudkowsky (2008a) notes that AI poses a special kind of existential risk, for it can surely destroy the human species but, if done right, it also has the unique capacity to *save* our species from all other existential risks. But will AI come before other existential risks, especially the risks of synthetic biology? How should efforts be allocated between safe AI and the mitigation of other existential risks? Is Oracle AI enough to mitigate other existential risks?

Which problems do we need to solve, and which ones can we have an AI solve? Can we get an AI to do Friendly AI philosophy *before* it takes over the world? Which problems must be solved by humans, and which ones can we hand off to the AI?

How can we develop microeconomic models of WBEs and self-improving systems? Hanson (1994, 1998, 2008a, 2008b, 2008c, forthcoming) provides some preliminary steps. Might such models help us predict takeoff speed and the likelihood of monopolar (singleton) vs. multipolar outcomes?.

The Noddy problem

An episode of the Noddy animated series has the following plot.

Noddy needs to go pick up Martha Monkey at the station. But it's such a nice, sunny day that he would prefer to play around outside. He gets an idea to solve this dilemma. He casts a duplication spell on himself and his car and tells the duplicate to go fetch Martha while he goes out to play. Later, Noddy is out having fun when he suddenly spots his duplicate. It turns out that the duplicate also preferred playing outside to doing the errand so he also cast a duplication spell. Then they see another duplicate, and another...

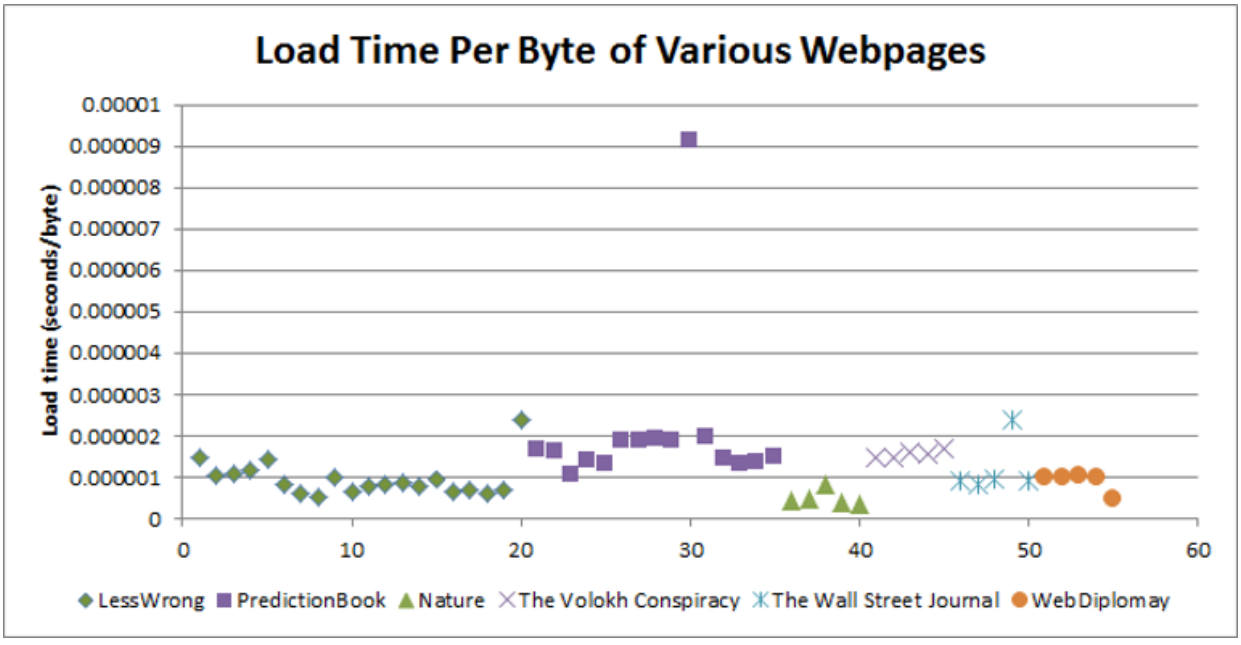
I think this story makes for a nice simple illustration of one of our perennial decision theoretic issues: When making decisions you should take into account that agents identical to yourself will make the same decision in the same situation. A common real-life example of the Noddy problem is when we try to pawn off our dietary problems to our future selves.

[META] My Negative Results

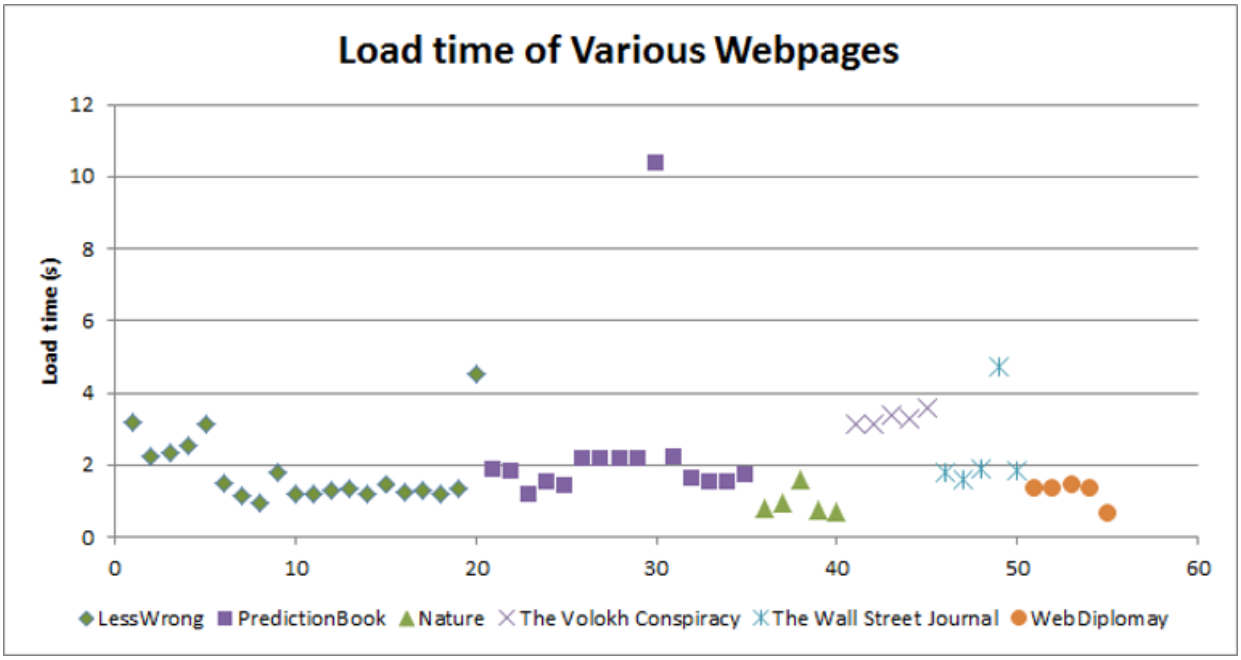
Yesterday, I made a post asking if anyone else had noticed LW being particularly slow. I offered to collect data on this, and was fairly sure (Probably about 80%) that it would show that LW loaded slower than other webpages. I took the post down after about 20 seconds (sorry if I confused you) since it was almost entirely insubstantial, and resolved to collect some actual data to report.

So I did that. Data was taken using this website. I was on my school's wireless network at the time, running Firefox 8.0. I didn't think to disable my addons before doing this, but I was running Adblock Plus, FastestFox, Greasemonkey, IE Tab 2, Movable Firefox Button, Omnibar, and Web of Trust. Data points were generated at the following websites. Each website was measured five times.
http://lesswrong.com/r/discussion/new
http://lesswrong.com
http://lesswrong.com/user/RobertLumley
http://lesswrong.com/promoted/
http://predictionbook.com
http://predictionbook.com/predictions
http://predictionbook.com/users/rlumley
http://www.nature.com/news/index.html
http://volokh.com/
http://online.wsj.com/public/page/news-opinion-commentary.html
http://www.webdiplomacy.net/index.php

After doing this, I downloaded an offline copy of each of these websites, and calculated load time per byte of website size. I plotted these results. To my surprise, LessWrong ended up being one of the fastest, although PredictionBook could use some work. I considered deleting the obvious outlier, but trends are clear even with it in there, and all things equal, I'd rather not delete data. Data points (in groups of 5) correspond to the webpages above, in order, ie. the first five points are from the discussion page of LW.



Surprisingly, LW is still one of the best even when only raw load time is compared. But the discussion page (remember, that's the first 5 points) takes somewhat longer than the other pages:



The excel spreadsheet used to generate this can be viewed here.

Since there is no real problem here, and it was all in my head, I considered not publishing this, but after all of the discussion we have about positive bias, I wasn't about to turn around and do the same thing. So here they are: my negative results.

The Human's Hidden Utility Function (Maybe)

Suppose it turned out that humans violate the axioms of VNM rationality (and therefore don't act like they have utility functions) because there are *three* valuation systems in the brain that make conflicting valuations, and all three systems contribute to choice. And suppose that upon reflection we would clearly reject the outputs of two of these systems, whereas the third system looks something more like a utility function we might be able to use in CEV.

What I just described is part of the leading theory of choice in the human brain.

Recall that human choices are made when certain populations of neurons encode expected subjective value (in their firing rates) for each option in the choice set, with the final choice being made by an argmax or reservation price mechanism.

Today's news is that our best current theory of human choices says that at least three *different* systems compute "values" that are then fed into the final choice circuit:

- The *model-based system* "uses experience in the environment to learn a model of the transition distribution, outcomes and motivationally-sensitive utilities." (See Sutton & Barto 1998 for the meanings of these terms in reinforcement learning theory.) The model-based system also "infers choices by... building and evaluating the search decision tree to work out the optimal course of action." In short, the model-based system is responsible for goal-directed behavior. However, making all choices with a goal-directed system using something like a utility function would be computationally prohibitive (Daw et al. 2005), so many animals (including humans) first evolved much simpler methods for calculating the subjective values of options (see below).

- The *model-free system* also learns a model of the transition distribution and outcomes from experience, but "it does so by caching and then recalling the results of experience rather than building and searching the tree of possibilities. Thus, the model-free controller does not even represent the outcomes... that underlie the utilities, and is therefore not in any position to change the estimate of its values if the motivational state changes. Consider, for instance, the case that after a subject has been taught to press a lever to get some cheese, the cheese is poisoned, so it is no longer worth eating. The model-free system would learn the utility of pressing the lever, but would not have the informational wherewithal to realize that this utility had changed when the cheese had been poisoned. Thus it would continue to insist upon pressing the lever. This is an example of motivational insensitivity."

- The *Pavlovian system*, in contrast, calculates values based on a set of hard-wired preparatory and consummatory "preferences." Rather than calculate value based on what is likely to lead to rewarding and punishing outcomes, the Pavlovian system calculates values consistent with automatic approach toward appetitive stimuli, and automatic withdrawal from aversive stimuli. Thus, "animals cannot help but approach (rather than run away from) a source of food, even if the experimenter has cruelly arranged things in a looking-glass world so

> that the approach appears to make the food recede, whereas retreating would make the food more accessible (Hershberger 1986)."

Or, as Jandila put it:

- *Model-based system*: Figure out what's going on, and what actions maximize returns, and do them.
- *Model-free system*: Do the thingy that worked before again!
- *Pavlovian system*: Avoid the unpleasant thing and go to the pleasant thing. Repeat as necessary.

In short:

> We have described three systems that are involved in making choices. Even in the case that they share a single, Platonic, utility function for outcomes, the choices they express can be quite different. The model-based controller comes closest to being Platonically appropriate... The choices of the model-free controller can depart from current utilities because it has learned or cached a set of values that may no longer be correct. Pavlovian choices, though determined over the course of evolution to be appropriate, can turn out to be instrumentally catastrophic in any given experimental domain...
>
> [Having multiple systems that calculate value] is [one way] of addressing the complexities mentioned, but can lead to clashes between Platonic utility and choice. Further, model-free and Pavlovian choices can themselves be inconsistent with their own utilities.

We don't yet know how choice results from the inputs of these three systems, nor how the systems might interact before they deliver their value calculations to the final choice circuit, nor whether the model-based system *really* uses anything like a coherent utility function. But it looks like the human *might* have a "hidden" utility function that would reveal itself if it wasn't also using the computationally cheaper model-free and Pavlovian systems to help determine choice.

At a glance, it seems that upon reflection I might embrace an extrapolation of the model-based system's preferences as representing "my values," and I would reject the outputs of the model-free and Pavlovian systems as the outputs of dumb systems that evolved for their computational simplicity, and can be seen as ways of trying to approximate the full power of a model-based system responsible for goal-directed behavior.

On the other hand, as Eliezer points out, perhaps we ought to be suspicious of this, because "it sounds like the correct answer ought to be to just keep the part with the coherent utility function in CEV which would make it way easier, but then someone's going to jump up and say: 'Ha ha! Love and friendship were actually in the other two!'"

Unfortunately, it's too early to tell whether these results will be useful for CEV. But it's a *little* promising. This is the kind of thing that sometimes happens when you hack away at the edges of hard problems. This is also a repeat of the lesson that "you can often out-pace most philosophers simply by reading what today's leading *scientists* have to say about a given topic instead of reading what *philosophers* say about it."

(For pointers to the relevant experimental data, and for an explanation of the mathematical role of each valuation system in the brain's reinforcement learning

system, see Dayan (2011). All quotes in this post are from that chapter, except for the last one.)

Introducing Leverage Research

Geoff Anders asked me to post this introduction to Leverage Research. Several friends of the Singularity Institute are now with Leverage Research, and we have overlapping goals.

> Hello Less Wrong! I'm Geoff Anders, founder of Leverage Research. Many Less Wrong readers are already familiar with Leverage. But many are not, and because of our ties to the Less Wrong community and our deep interest in rationality, I thought it would be good to formally introduce ourselves.
>
> I founded Leverage at the beginning of 2011. At that time we had six members. Now we have a team of more than twenty. Over half of our people come from the Less Wrong / Singularity Institute community. One of our members is Jasen Murray, the leader of the Singularity Institute's recent Rationality Boot Camp. Another is Justin Shovelain, a two-year Visiting Fellow at SIAI and the former leader of their intelligence amplification research. A third is Adam Widmer, a former co-organizer of the New York Less Wrong group.
>
> Our goal at Leverage is to make the world a much better place, using the most effective means we can. So far, our conclusion has been that the most effective way to change the world is by means of high-value projects, projects that will have extremely positive effects if they succeed and that have at least a fair probability of success.
>
> One of our projects is existential risk reduction. We have conducted a study of the efficacy of methods for persuading people to take the risks of artificial general intelligence (AGI) seriously. We have begun a detailed analysis of AGI catastrophe scenarios. We are working with risk analysts inside and outside of academia. Ultimately, we intend to achieve a comprehensive understanding of AGI and other global risks, develop response plans, and then enact those plans.
>
> A second project is intelligence amplification. We have reviewed the existing research and analyzed current approaches. We then created an initial list of research priorities, ranking techniques by likelihood of success, likely size of effect, safety, cost and so on. We plan to start testing novel techniques soon.
>
> These are just two of our projects. We have several others, including the development of rationality training program, the construction and testing of theories of the human mind and an investigation of the laws of idea propagation.
>
> Changing the world is a complex task. Thus we have a plan that guides our efforts. We know that to succeed, we need to become better than we are. So we take training and self-improvement very seriously. Finally, we know that to succeed, we need more talented people. If you want to significantly improve the world, are serious about self-improvement and believe that changing the world means we need to work together, contact us. We're looking for people who are interested in our current projects or who have ideas of their own.
>
> We've been around for just over a year. In that time we've gotten many of our projects underway. We doubled once in our first six months and again in our

second six months. And we have just set up our first physical location, in New York City.

If you want to learn more, visit our website. If you want to get involved, want to send a word of encouragement, or if you have suggestions for how we can improve, write to us.

With hope for the future,

Geoff Anders, on behalf of the Leverage Team

Shit Rationalists Say?

I assume everyone has run across at least one of the "Shit X's Say" format of videos? Such as Shit Skeptics Say. When done right it totally triggers the in-group warm-fuzzies. (Not to be confused with the nearly-identically formatted "Shit X's Say to Y's" which is mainly a way for Y's to complain about X's).

What sort of things do Rationalists often say that triggers this sort of in-group recognition which could be popped into a short video? A few I can think of...

You should sign up for cryonics. I want to see you in the future.

…intelligence explosion…

What's your confidence interval?

You know what they say: one man's Modus Ponens is another man's Modus Tollens

This may sound a bit crazy right now, but hear me out…

What are your priors?

When the singularity comes that won't be a problem anymore.

I like to think I'd do that, but I don't fully trust myself. I am running on corrupted hardware after all.

I want to be with you, and I don't foresee that changing in the near future.

…Bayesian statistics…

So Omega appears in front of you…

What would you say the probability of that event is, if your beliefs are true?

Others?

"Politics is the mind-killer" is the mind-killer

Summary: I propose we somewhat relax our stance on political speech on Less Wrong.

Related: The mind-killer, Mind-killer

A recent series of posts by a well-meaning troll (example) has caused me to re-examine our "no-politics" norm. I believe there has been some unintentional creep from the original intent of Politics is the Mind-Killer. In that article, Eliezer is arguing that discussions here (actually on *Overcoming Bias*) should not use examples from politics in discussions that are not about politics, since they distract from the lesson. Note the final paragraph:

> I'm not saying that I think *Overcoming Bias* should be apolitical, or even that we should adopt Wikipedia's ideal of the Neutral Point of View. But try to resist getting in those good, solid digs if you can possibly avoid it. If your topic legitimately relates to attempts to ban evolution in school curricula, then go ahead and talk about it - but don't blame it explicitly on the whole Republican Party; some of your readers may be Republicans, and they may feel that the problem is a few rogues, not the entire party. As with Wikipedia's NPOV, it doesn't matter whether (you think) the Republican Party really *is* at fault. It's just better for the spiritual growth of the community to discuss the issue without invoking color politics.

So, the original intent was not to ban political speech altogether, but to encourage us to come up with less-charged examples where possible. If the subject you're really talking about is politics, and it relates directly to rationality, then you should be able to post about it without getting downvotes strictly because "politics is the mind-killer".

It could be that this drift is less of a community norm than I perceive, and there are just a few folks (myself included) that have taken the original message too far. If so, consider this a message just to those folks such as myself.

Of course, politics would *still* be off-topic in the comment threads of most posts. There should probably be a special open thread (or another forum) to which drive-by political activists can be directed, instead of simply saying "We don't talk about politics here".

David_Gerard makes a similar point here (though FWIW, I came up with this title independently).

Leveling Up in Rationality: A Personal Journey

See also: Reflections on rationality a year out

My favorite part of *Lord of the Rings* was skipped in both film adaptations. It occurs when our four hobbit heroes (Sam, Frodo, Merry and Pippin) return to the Shire and learn it has been taken over by a gang of ruffians. Merry assumes Gandalf will help them free their home, but Gandalf declines:

> I am not coming to the Shire. You must settle its affairs yourselves; that is what you have been trained for... My dear friends, you will need no help. You are grown up now. Grown indeed very high...

As it turns out, the hobbits *have* acquired many powers along their journey — powers they use to lead a resistance and free the Shire.

That is how I felt when I flew home for the holidays this December. Minnesota wasn't ruled by ruffians, but the familiar faces and places reminded me of the person I had been before I moved away, just a few years ago.

And I'm just *so* much more powerful than I used to be.

And in *my* case, at least, many of my newfound powers seem to come from having seriously leveled up in *rationality*.

Power 0: Curiosity

I was always "curious," by which I mean I *felt* like I wanted to know things. I read lots of books and asked lots of questions. But I didn't *really* want to know the truth, because I didn't care enough about the truth to study, say, probability theory and the cognitive science of how we deceive ourselves. I just studied different Christian theologies — and, when I was *really* daring, different supernatural religions — and told myself *that* was what honest truth-seeking looked like.

It took *20 years* for reality to pierce my comfortable, carefully cultivated bubble of Christian indoctrination. But when it finally popped, I realized I had (mostly) wasted my life thus far, and I was *angry*. Now I studied things not just for the pleasure of discovery and the gratifying *feeling* of caring about truth, but because I *really* wanted an accurate model of the world so I wouldn't do stupid things like waste two decades of life.

And it was this curiosity, more than anything else, that led to everything else. So long as I burned for reality, I was bound to level up.

Power 1: Belief Propagation

One factor that helped religion cling to me for so long was my ability to compartmentalize, to shield certain parts of my beliefs from attack, to apply different standards to different beliefs like the scientist outside the laboratory. When genuine curiosity tore down those walls, it didn't take long for the implications of my atheism to propagate. I noticed that contra-causal free will made no sense for the same reasons God made no sense. I noticed that whatever value existed in the universe was made of atoms. I assumed the basics of transhumanism without knowing there was a thing called "transhumanism." I noticed that minds didn't need to be made of meat, and that machines could be made more moral than humans. (I called them "artificial superbrains" at the time.) I noticed that scientific progress could actually be *bad*, because it's easier to destroy the world than to protect it. I also noticed we should therefore "encourage scientific research that saves and protects lives, and discourage scientific research that may destroy us" — and this was *before* I had read about existential risk and "differential technological development."

Somehow, I didn't notice that naturalism + scientific progress also implied intelligence explosion. I had to *read* that one. But when I did, it set off another round of rapid belief updates. I noticed that the entire world could be lost, that moral theory was an urgent engineering problem, that technological utopia is actually possible (however unlikely), and more.

The power of belief propagation gives me clarity of thought and coherence of action. My actions are now less likely to be informed by multiple incompatible beliefs, though this still occurs *sometimes* due to cached thoughts.

Power 2: Scholarship

I was always one to look things up, but before my deconversion my scholarship heuristic seems to have been "Find something that shares most of my assumptions and tells me roughly what I want to hear, filled with lots of evidence to reassure me of my opinion." That's not what I *thought* I was doing at the time, but looking back at my reading choices, that's what it *looks* like I was doing.

After being taken by genuine curiosity, my heuristic became something more like "Check what the mainstream scientific consensus is on the subject, along with the major alternative views and most common criticisms." Later, I added qualifications like "But watch out for signs that an entire field of inquiry is fundamentally unsound."

The power of *looking shit up* proved to have enormous practical value. How could I make *Common Sense Atheism* popular, quickly? I studied how to build blog traffic, applied the major lessons, and within 6 months I had one of the most popular atheism blogs on the internet. How could I improve my success with women? I skim-read dozens of books on the subject, filtered out the best advice, applied it (after much trepidation), and eventually had enough success that I didn't need to worry about it anymore. What are values, and how do they work? My search lead me from philosophy to affective neuroscience and finally to neuroeconomics, where I hit the jackpot and wrote A Crash Course in the Neuroscience of Human Motivation. How could I be happier? I studied the science of happiness, applied its lessons, and went from occasionally suicidal to stably happy. How could I make the Singularity Institute more effective? I studied non-profit management and fundraising, and am currently (with lots of help) doing quite a lot to make the organization more efficient and credible.

My most useful scholarship win had to do with beating akrasia. Eliezer wrote a post about procrastination that drew from personal anecdote but not a single experiment. This prompted me to write my first post, which suggested he ought to have done a bit of research on procrastination, so he could stand on the shoulders of giants. A simple Google scholar search on "procrastination" turned up a recent "meta-analytic and theoretical review" of the field as the 8th result, which pointed me to the resources I used to write How to Beat Procrastination. Mastering that post's algorithm for beating akrasia might be the most useful thing I've ever done, since it empowers everything else I try to do.

Power 3: Acting on Ideas

Another lesson from my religious deconversion was that *abstract ideas have consequences*. Because of my belief in the supernatural, I had spent 20 years (1) studying theology instead of math and science, (2) avoiding sexual relationships, and (3) training myself in fantasy-world "skills" like prayer and "sensing the Holy Spirit." If I wanted to benefit from having a more *accurate* model of the world as much as I had been harmed by having a false model, I'd need to actually *act* in response to the most probable models of the world I could construct.

Thus, when I realized I didn't like the Minnesota cold and could be happy without seeing my friends and family that often, I threw all my belongings in my car and moved to California. When I came to take intelligence explosion seriously, I quit my job in L.A., moved to Berkeley, interned with the Singularity Institute, worked hard, got hired as a researcher, and was later appointed Executive Director.

Winning with Rationality

These are just a few of my rationality-powers. Yes, I could have gotten these powers another way, but in my case they seemed to flow largely from that first virtue of rationality: genuine curiosity. Yes, I've compressed my story and made it sound less messy than it really was, but I do believe I've been gaining in rationalist power — the power of agency, systematized winning — and that my life is much better as a result. And yes, *most* people won't get these results, due to things like akrasia, but maybe if we figure out how to teach the unteachable, those chains won't hold us anymore.

What does a Level 60 rationalist look like? Maybe Eliezer Yudkowsky + Tim Ferris? That sounds like a worthy goal! A few dozen people *that* powerful might be able to, like, save the world or something.

What Curiosity Looks Like

See also: Twelve Virtues of Rationality, The Meditation on Curiosity, Use Curiosity

What would it look like if someone was *truly curious* — if they *actually wanted true beliefs*? Not someone who wanted to *feel* like they sought the truth, or to *feel* their beliefs were justified. Not someone who wanted to signal a desire for true beliefs. No: someone who *really* wanted true beliefs. What would that look like?

A truly curious person would seek to understand the world as broadly and deeply as possible. They would study the humanities but especially math and the sciences. They would study logic, probability theory, argument, scientific method, and other core tools of truth-seeking. They would inquire into epistemology, the study of knowing. They would study artificial intelligence to learn the algorithms, the *math*, the laws of how an *ideal* agent would acquire true beliefs. They would study modern psychology and neuroscience to learn how their brain acquires beliefs, and how those processes depart from ideal truth-seeking processes. And they would study how to minimize their thinking errors.

They would practice truth-seeking skills as a musician practices playing her instrument. They would practice "debiasing" techniques for reducing common thinking errors. They would seek out contexts known to make truth-seeking more successful. They would ask others to help them on their journey. They would ask to be held accountable.

They would cultivate that burning itch to *know*. They would admit their ignorance but seek to destroy it.

They would be precise, not vague. They would be clear, not obscurantist.

They would not flinch away from experiences that might destroy their beliefs. They would train their emotions to fit the facts.

They would update their beliefs quickly. They would resist the human impulse to rationalize.

But even *all this* could merely be a signaling game to increase their status in a group that rewards the appearance of curiosity. Thus, the final test for genuine curiosity is *behavioral change*. You would find a genuinely curious person studying and learning. You would find them practicing the skills of truth-seeking. You wouldn't merely find them saying, "Okay, I'm updating my belief about that" — you would also find them making decisions consistent with their new belief and inconsistent with their former belief.

Every week I talk to people who say they are trying to figure out the truth about something. When I ask them a few questions about it, I often learn that they know almost nothing of logic, probability theory, argument, scientific method, epistemology, artificial intelligence, human cognitive science, or debiasing techniques. They do not regularly practice the skills of truth-seeking. They don't seem to say "oops" very often, and they change their behavior even *less* often. I conclude that they probably want to *feel* they are truth-seeking, or they want to *signal* a desire for truth-seeking, or they might even self-deceivingly "believe" that they place a high value on knowing the truth. But their actions show that they aren't trying very hard to have true beliefs.

Dare I say it? Few people *look* like they really want true beliefs.

The Personality of (great/creative) Scientists: Open and Conscientious

We've discussed the Big Five in the past, such as the relationship of Openness to parasites & signaling or whether hallucinogens increase Openness and parasites decrease it, along with my little notes on the value of Conscientiousness. This is another entry in the topic of 'what is Big Five good for'.

I researched the topic of how and whether Conscientiousness and Openness correlate with scientific achievement for Luke for the Intelligence Explosion paper; here is some of what I found:

1. "Creativity, Intelligence, and Personality", 1981 review:

   > "Studies of creative adult artists, scientists, mathematicians, and writers find them scoring very high on tests of general intelligence (e.g. Barron 1969; Bachtold & Werner 1970; Helson & Crutchfield 1970b; Cattell 1971; Helson 1971; Bachtold & Werner 1973; Gough 1976a), though rs between tested intelligence and creative achievement in these samples range from insignificantly negative (r = -.05, Gough 1976a) to mildly and significantly positive (r = +.31, Helson 1971)."

   I found this one amusing:

   > "It should be noted that creative people are often perceived and rated as more intelligent than less creative people even in samples where no corresponding correlations between tested intelligence and creativity obtain. Despite an r of -.08 between Terman's Concept Mastery Test and professionally rated creativity among the top 40 IPAR architects (MacKinnon 1962a), e.g., staff ratings of the single adjective "intelligent" correlated +.39 with the index of creativity (MacKinnon 1966). While such an r may reflect some spurious halo effects, it may also tell us something about the true overlap in meaning of these terms in the natural language."

   An embarrassment of riches; no summary, but a starting point if one needs more:

   > "*SCIENCE AND TECHNOLOGY*: Personality correlates of scientific achievement and creativity were studied in elementary school children (Milgram et al 1977); high school students (Schaefer & Anastasi 1968, Parloff et al 1968, Anastasi & Schaefer 1969, Schaefer 1969a, b, Walberg 1969a); undergraduates, young adults, and graduate students (Rossman & Horn 1972, Schaefer 1973, Gough 1979, Korb & Frankiewicz 1979); psychologists (Chambers 1964, Wispe 1965, Bachtold & Werner 1970); inventors (Bergum 1975, Albaum 1976, Albaum & Baker 1977); mathematicians (Helson 1967b, 1968a, Parloff et al 1968; Helson & Crutchfield 1970a, b; Helson 1971; Gough 1979); chemists (Chambers 1964); and assorted engineers and research scientists (McDermid 1965, Owens 1969, Bachtold & Werner 1972, Bergum 1973, Eiduson 1974, Gough 1979)."

2. "How development and personality influence scientific thought, interest, and achievement", GJ Feist, *Review of General Psychology*, 2006; important bits start

on pg 9:

"In 1998, I published a quantitative review of the literature on personality and scientific interest and creativity (Feist, 1998. In this meta-analytic review of which personality traits make interest and creativity in science more likely, I found every published (and some unpublished studies) that examined the role in personality in scientific interest or scientific creativity from 1950 to 1998. There were 26 studies that reported quantitative effects of personality in scientists compared to non-scientists.

...The two strongest effect sizes (medium in magnitude) were for the positive and negative poles of conscientiousness (C; see Table 1). Being high in conscientiousness (C_2) consists of scales and items such as careful, cautious, conscientious, fastidious, and self-controlled, whereas being low in conscientiousness (C_X) consists only of two scales/items, namely, direct expression of needs and psychopathic deviance. Although the C_X dimension comprised only five comparisons, it is clear that relative to non-scientists, scientists are roughly a half a standard deviation higher on conscientiousness and controlling of impulses. In addition, low openness to experience had a median d of .30, whereas introversion had a median effect size of .26.

A consistent finding in the personality and creativity in science literature has been that creative and eminent scientists tend to be more open to experience and more flexible in thought than less creative and eminent scientists (see Table 2). Many of these findings stem from data on the flexibility (Fe) and tolerance (To) scales of the California Psychological Inventory (Feist & Barron, 2003; Garwood, 1964; Gough, 1961; Helson, 1971; Helson & Crutchfield, 1970; Parloff & Datta, 1965). The Fe scale, for instance, taps into flexibility and adaptability of thought and behavior as well as the preference for change and novelty (Gough, 1987). The few studies that have reported either no effect or a negative effect of flexibility in scientific creativity have been with student samples (Davids, 1968; Smithers & Batcock, 1970).

For instance, Feist and Barron (2003) examined personality, intellect, potential, and creative achievement in a 44-year longitudinal study. More specifically, they predicted that personality would explain unique variance in creativity over and above that already explained by intellect and potential. Results showed that observer-rated Potential and Intellect at age 27 predicted Lifetime Creativity at age 72, and yet personality variables (such as Tolerance and Psychological Mindedness) explained up to 20% of the variance over and above Potential and Intellect. Specifically, two measures of personality—California Psychological Inventory scales of Tolerance (To) and Psychological Mindedness (Py)–resulted in the 20% increase in variance explained (20%) over and above potential and intellect. The more tolerant and psychologically minded the student was, the more likely he was to make creative achievements over his lifetime. Together, the four predictors (Potential, Intellect, Tolerance, and Psychological Mindedness) explained a little more than a third of the variance in lifetime creative achievement. I should point out that these findings on To and Py mirror very closely those reported by Helson and Pals (2000) in a longitudinal study of women from age 21 to 52.

> …Busse and Mansfield (1984), for example, studied the personality characteristics of 196 biologists, 201 chemists, and 171 physicists, and commitment to work (i.e., "need to concentrate intensively over long periods of time on one's work") was the strongest predictor of productivity (i.e., publication quantity) even when holding age and professional age constant. Helmreich, Spence, Beane, Lucker, and Matthews (1980) studied a group of 196 academic psychologists and found that different components of achievement and drive had different relationships with objective measures of attainment (i.e., publications and citations). With a self-report measure, they assessed three different aspects of achievement: "mastery" preferring challenging and difficult tasks; "work" enjoying working hard; and "competitiveness" liking interpersonal competition and bettering others….Helmreich and his colleagues found that mastery and work were positively related to both publication and citation totals, whereas competitiveness was positively related to publications but negatively related to citations
>
> …Helson (1971) compared creative female mathematicians with less creative female mathematicians, matched on IQ. Observers blindly rated the former as having more "unconventional thought processes," as being more "rebellious and non-conforming," and as being less likely to judge "self and others in conventional terms." More recently, Rushton, Murray, and Paunonen (1987) conducted factor analyses of the personality traits most strongly loading on the "research" factor (in contrast to a "teaching" factor) in two separate samples of academic psychologists. Among other results, they found that "independence" tended to load on the research factor, whereas "extraversion" tended to load on the teaching factor."

Pretty substantial. The 26 study meta-analysis was Feist, G. J. (1998). "A meta-analysis of the impact of personality on scientific and artistic creativity". *Personality and Social Psychological Review*, 2, 290–309

Feist 1998 was also cited in a chapter of *The Cambridge Handbook of Creativity*, "the relationship between creativity and intelligence", but I couldn't get the book; further reading, if anyone wants some.

One useful bit from Feist; I also ran into a lot of CP Benbow-keyworded studies of the gifted SMPY cohorts to the effect that the kids' early interest in science predicts later careers in science, which would tie in nicely to this:

> "The empirical consensus is that early levels of high productivity do regularly predict continued levels of high productivity across one's lifetime (Cole, 1979; Dennis, 1966; Helson & Crutchfield, 1970; Horner et al., 1986; Lehman, 1953; Over, 1982; Reskin, 1977; Roe, 1965; Simonton, 1988a, 1991)….In other words, the younger NAS members were when they and others recognized their scientific talent, when they wanted to be a scientist, and when they first conducted scientific research, the younger they were when they published their first paper. Age of first publication in turn predicted total publication rate over the lifetime, meaning that the earlier one publishes, the more productive one will be. This pattern of relationships —from precocity to age of first publication to lifetime productivity—implies an indirect connection between precocity and publication rate. The only precocity variable that reached the .05 level of significance with lifetime productivity was age that one first conducted formal research."

3. "Scientific talent, training, and performance: Intellect, personality, and genetic endowment", Simonton 2008; the abstract caught my eye:

> "After specifying the ideal data requirements for the application of the three estimators, the procedures were applied to previously published results. Personality traits were illustrated with the use of the California Psychological Inventory and the Eysenck Personality Questionnaire with respect to two criteria (scientists versus non-scientists and creative scientists versus less creative scientists) and intellectual traits with the use of the Miller Analogies Test with respect to seven criteria (graduate grade-point average, faculty ratings, comprehensive examination scores, degree attainment, research productivity, etc.). The outcome provides approximate, *lower-bound estimates of the genetic contribution to scientific training and performance*.
>
> …Sawyer reviewed three investigations that allegedly disconfirm the role of genetic endowment in any form of creative achievement (viz., Barron, 1972; Reznikoff, Domino, Bridges, & Honeyman, 1973; Vandenberg, Stafford, & Brown, 1968).2 Likewise, Ericsson, Roring, and Nandagopal (2007) cited two behavior genetic investigations in drawing the same conclusion about talent in general (viz., Bouchard & Lykken, 1999; Klissouras et al., 2001).
>
> …This argument certainly applies to scientific talent. For instance, a comprehensive longitudinal study of the mathematically precocious (Lubinski, Webb, Morelock, & Benbow, 2001) has extremely few twins in the sample (D. Lubinski, personal communication, March 15, 2007). Simonton's (1991a) study of 2,026 eminent scientists contained only one twin (Auguste Picard). And, needless to say, there are no twins, monozygotic or dizygotic, among Nobel laureates in the sciences. Thus, not only may we lack direct evidence for scientific talent, but also it may never be possible to establish such substantiation with the use of standard behavior genetic methods.
>
> …Bouchard and Lykken (1999) demonstrated that the personality characteristics associated with scientific productivity display heritabilities ranging between .32 and .57, meaning that between 32% and 57% of the variance in those traits can be attributed to genetic endowment. Similarly, the Creativity Personality Scale (CPS) of the Adjective Check List (ACL) not only predicts scientific creativity (Gough, 1979) but also has a heritability of .54 (Bouchard & Lykken, 1999; Waller, Bouchard, Lykken, Tellegen, & Blacker, 1993). Hence, 54% of the variance in the CPS has a genetic contribution. Moreover, because predictive validities are known for this measure, we can draw a more powerful inference. For instance, CPS scores correlated .31 with the creativity ratings that expert judges assigned 57 mathematicians (Gough, 1979). This signifies that almost 10% of the variance in that criterion can be attributed to CPS (i.e., .312 6 .096). Multiplying the squared criterion–trait correlation by the heritability coefficient then yields .052, which implies that over 5% of the variance in the rated creativity of these mathematicians might be ascribed to the genetic part of the CPS scores.

"Bouchard & Lykken 1999" = Bouchard, T. J., Jr., & Lykken, D. T. (1999). "Genetic and environmental influence on correlates of creativity". In N. Colangelo & S. G. Assouline (Eds.), *Talent development III: Proceedings from the 1995 Henry B. & Jocelyn Wallace National Symposium on Talent Development* (pp. 81–97). I

couldn't find this, but I did find http://cogprints.org/611/1/genius.html which describes it a little more (C-f 'in press').

Moving onwards, pg 9–12 discuss Feist 1998's meta-analysis:

> …Even worse, "science" was obliged to include the physical, biological, and social sciences as well as mathematics, engineering, and invention. To the extent that the personality profiles are closely tailored to domain-specific training or performance criteria, this definitional inclusiveness implies that the hc2 estimates are too low. This criticism is not intended to fault Feist's (1998) meta-analytic review. To obtain sound effect size estimates he had no other option but to collate many diverse findings. Nevertheless, in this analysis of intellectual traits it is feasible to substitute somewhat more specific criteria for these more global contrasts.
>
> For example, spatial ability has been identified as a crucial component of math–science talent that exhibits predictive utility beyond that provided by both mathematical and verbal ability (Webb, Lubinski, & Benbow, 2007). Yet measures of spatial intelligence display heritabilities almost as high as general intelligence (Bratko, 1996; McClearn, Johansson, Berg, & Pedersen, 1997). Moreover, these more specialized intellectual abilities may be especially useful in differentiating distinct types of scientific talents. For instance, in Roe's (1953) classic study of 64 eminent scientists it was found that theoretical physicists, experimental physicists, biologists, psychologists, and anthropologists display distinctive profiles with respect to verbal, mathematical, and spatial intelligence.

Concluding:

> …It is not possible to say exactly when the final sum would be, but a conservative guess might be that between 10% and 20% of the variance in these criteria could be potentially attributed to genetic effects (cf. Simonton, 2007).
>
> For the sake of this discussion, then, suppose that .10 ∋ hc2 ∋ .20 holds for the training and performance criteria examined here. Does this outcome imply that scientific talent is an important substantive phenomenon? To answer this question requires that we obtain some kind of baseline for comparison.…[the range of estimates] can be qualitatively expressed as medium to large (Cohen, 1988). This range is about as good as can be expected of most effects in the behavioral sciences (Meyer et al., 2001; Rosenthal, 1990). To offer specific comparisons, the lower-end estimate is about the same magnitude as the relation between psychotherapy and subsequent well-being, whereas the upper-end estimate is about the same size as the correlation between height and weight among U.S. adults (Meyer et al., 2001).

[META] 'Rational' vs 'Optimized'

A new arrival, Kouran, recently challenged our conventional use of the label "rational" to describe various systems. The full thread is here, and it doesn't summarize neatly, but he observes that we often use "rational" in the context of non-intellectual, non-cognitive, etc. systems, and that this is an unconventional use of the word.

Unsurprisingly, this led to Standard Conversation Number 12 about how we don't really use "rational" to mean what the rest of the world means by it, and about instrumental rationality, and etc. and etc. In the course of that discussion I made the observation a couple of times (here and here) that we could probably substitute some form of "optimal" for "rational" wherever it appears without losing any information.

Of course, status quo bias being what it is, I promptly added that we wouldn't actually want to do that, because, y'know, it would be work and involve changing stuff.

But the more I think about it, the more it seems like I *ought* to endorse that lexical shift. We do spend a not-inconsiderable amount of time and attention on alleviating undesirable side-effects of the word 'rational,' such as the Spock effect, and our occasional annoying tendency to talk about the 'rational' choice of shoe-polish when we really mean the optimal choice, and our occasional tendency to tie ourselves in knots around "rationalists should *win*". (That optimized systems do better than non-optimized systems is pretty much the *definition* of "optimized," after all. If we say that rational systems generally do better than irrational systems, we're saying that rational systems are generally optimal, which is a non-empty statement. But if we define "rational" to mean the thing that wins, which we sometimes do, it seems simpler to talk about optimized systems in the first place.)

There's precedent for this... a while ago I started getting out of the habit of talking about "artificial intelligences" when I really wanted to talk about superhuman optimizing systems instead, and I continue to endorse that change. So, I'm going to stop using "rational" when I actually mean optimal. I encourage others to do so as well. (Or, conversely, to tell me why I shouldn't.)

This should go without saying, but in case it doesn't: I'm not proposing recoding anything or rewriting anything or doing any work here beyond changing my use of language as it's convenient for me to do so.