Best of LessWrong: February 2013

1. An attempt to dissolve subjective expectation and personal identity
2. Philosophical Landmines
3. Imitation is the Sincerest Form of Argument
4. Pinpointing Utility
5. A brief history of ethically concerned scientists
6. Rationalist Lent
7. Official LW uncensored thread (on Reddit)
8. The Singularity Wars
9. Three Axes of Prohibitions
10. The Fundamental Question - Rationality computer game design
11. Visual Mental Imagery Training
12. Naturalism versus unbounded (or unmaximisable) utility options
13. Learning critical thinking: a personal example
14. Memetic Tribalism

Best of LessWrong: February 2013

1. An attempt to dissolve subjective expectation and personal identity
2. Philosophical Landmines
3. Imitation is the Sincerest Form of Argument
4. Pinpointing Utility
5. A brief history of ethically concerned scientists
6. Rationalist Lent
7. Official LW uncensored thread (on Reddit)
8. The Singularity Wars
9. Three Axes of Prohibitions
10. The Fundamental Question - Rationality computer game design
11. Visual Mental Imagery Training
12. Naturalism versus unbounded (or unmaximisable) utility options
13. Learning critical thinking: a personal example
14. Memetic Tribalism

An attempt to dissolve subjective expectation and personal identity

> *I attempt to figure out a way to dissolve the concepts of 'personal identity' and 'subjective expectation' down to the level of cognitive algorithms, in a way that would let one bite the bullets of the anthropic trilemma. I proceed by considering four clues which seem important: 1) the evolutionary function of personal identity, 2) a sense of personal identity being really sticky, 3) an undefined personal identity causing undefined behavior in our decision-making machinery, and 4) our decision-making machinery being more strongly grounded in our subjective expectation than in abstract models. Taken together, these seem to suggest a solution.*

I ended up re-reading some of the debates about the anthropic trilemma, and it struck me odd that, aside for a few references to personal identity being an evolutionary adaptation, there seemed to be no attempt to reduce the concept to the level of cognitive algorithms. Several commenters thought that there wasn't really any problem, and Eliezer asked them to explain why the claim of there not being any problem regardless violated the intuitive rules of subjective expectation. That seemed like a very strong indication that the question needs to be dissolved, but almost none of the attempted answers seemed to do that, instead trying to solve the question via decision theory without ever addressing the core issue of subjective expectation. rwallace's I-less Eye argued - I believe correctly - that subjective anticipation isn't ontologically fundamental, but still didn't address the question of why it feels like it is.

Here's a sketch of a dissolvement. It seems relatively convincing to me, but I'm not sure how others will take it, so let's give it a shot. Even if others find it incomplete, it should at least help provide clues that point towards a better dissolvement.

Clue 1: The evolutionary function of personal identity.

Let's first consider the *evolutionary* function. Why have we *evolved* a sense of personal identity?

The first answer that always comes to everyone's mind is that our brains have evolved for the task of spreading our genes, which involves surviving at least for as long as it takes to reproduce. Simpler neural functions, like maintaining a pulse and having reflexes, obviously do fine without a concept of personal identity. But if we wish to use abstract, explicit reasoning to advance our own interests, we need some definition for exactly *whose* interests it is that our reasoning process is supposed to be optimizing. So evolution comes up with a fuzzy sense of personal identity, so that optimizing the interests of this identity also happens to optimize the interests of the organism in question.

That's simple enough, and this point was already made in the discussions so far. But that doesn't feel like it would resolve our confusion yet, so we need to look at the way that personal identity is actually implemented in our brains. What is the *cognitive* function of personal identity?

Clue 2: A sense of personal identity is really sticky.

Even people who disbelieve in personal identity don't really seem to disalieve it: for the most part, they're just as likely to be nervous about their future as anyone else. Even advanced meditators who go out trying to dissolve their personal identity seem to still retain some form of it. PyryP claims that at one point, he reached a stage in meditation where the experience of "somebody who experiences things" shattered and he could turn it entirely off, or attach it to something entirely different, such as a nearby flower vase. But then the experience of having a self began to come back: it was as if the brain was hardwired to maintain one, and to reconstruct it whenever it was broken. I asked him to comment on that for this post, and he provided the following:

> It seems like my consciousness is rebuilding a new ego on top of everything, one which is not directly based on feeling as one with a physical body and memories, but which still feels like it is the thing that experiences whatever happens.
>
> To elaborate, many things in life affect the survival and success of an organism. Even though the organism would never experience itself as being separate from the surrounding universe, in ordinary life it's still useful to have concepts relating to the property and values of the organism. But even this pragmatic approach is enough for the ego-construction machinery, and the same old bad habits start to stick on it, even though the organism doesn't have any experience of itself that would be compatible with having a persistent 'soul'.
>
> Habits probably don't stick as strongly as they did before seeing the self as an illusion, but I'm still the same old asshole in certain respects. That might have something to do with the fact that I have no particular need to be particularly holy and clean. The ego-construction process is starting to be sufficiently strong that I've even began doubting how big of a change this has been. I don't have a clear recollection of whether I feel considerably different now than before, anymore.
>
> I still think that the change I experienced was a positive one and I feel like I can enjoy life in a less clinging way. I don't know if I've gained any special talents regarding my outlook of life that couldn't be maintained with simple mindfullness. I however do experience this certain transpersonal flow that makes everything lighter and easier. Something that makes the basic mindfullness effortless. I may also be making this shit up. The sunk cost of having spent lots of time in meditation makes people say funny things about their achievements. There is this insisting feeling that something is different, dammit.
>
> Anyway meditation is great fun and you can get all kinds of extremely pleasurable experiences with it. Don't read that if you're having trouble with desire getting in the way of your meditation. Ooops. Should've putten these the other way around. Yeah. I'm a dick. Deal with it. :P

Also, I know him in real life, and he doesn't really come off as behaving all that differently from anybody else.

Then there's also the fact that we seem to be almost incapable of thinking in a way that wouldn't still implicitly assume some concept of personal identity behind it. For example, I've said things like "it's convenient for me to disbelieve in personal identity at times, because then the person now isn't the same one as the person tomorrow, so I don't need to feel nervous about what happens to the person tomorrow". But here I'm not actually disbelieving in personal identity – after all, I clearly believe that there exists some "is-a" type relation that I can use to compare myself today and myself

tomorrow, and which returns a negative. If I truly disbelieved in personal identity, I wouldn't even have such a relation: asking “is the person today the same as the person tomorrow” would just return *undefined*.

Clue 3: Our decision-making machinery exhibits undefined behavior in the presence of an undefined personal identity.

This seems like an important thing to notice. What would it imply if I really didn't have *any* concept of personal identity or subjective expectation? If I asked myself whether I'd be the same person tomorrow as I was today, got an *undefined* back, and tried to give that as input to the systems actually driving my behavior... what would they say I should do?

Well, I guess it would depend on what those systems valued. If I was a paperclipper running on a pure utility-maximizing architecture, I guess they might say “who cares about personal identity anyway? Let's make paperclips!”.

But in fact, I'm a human, which means that a large part of the algorithms that actually drive my behavior are defined by reference to a concept of personal identity. So I'd ask them “I want to play computer games but in reality I should really study instead, which one do I actually do?”, and they'd reply “well let's see, to answer that we'd need to consider that which you expect to experience in the short term versus that which you expect to experience in the long term... AIEEEEEEE NULL POINTER EXCEPTION” and then the whole system would crash and need to be rebooted.

Except that it wouldn't, because it has been historically rather hard to reboot human brains, so they've evolved to handle problematic contingencies in other ways. So what probably would happen is that the answer would be “umm, we don't know, give us a while to work that out” and then some other system that didn't need a sense of identity to operate would take over. We'd default to some old habit, perhaps. In the meanwhile, the brain would be regenerating a concept of personal identity in order to answer the orignal question and things would go back to normal. And as far as I can tell, that's actually roughly what seems to happen.

Eliezer asked:

> It seems to me that there's some level on which, even if I say very firmly, "I now resolve to care only about future versions of myself who win the lottery! Only those people are defined as Eliezer Yudkowskys!", and plan only for futures where I win the lottery, then, come the next day, I wake up, look at the losing numbers, and say, "Damnit! What went wrong? I thought personal continuity was strictly subjective, and I could redefine it however I wanted!"

One possible answer could be that even if Eliezer did succeed in reprogramming his mind to think in such a weird, unnatural way, that would leave the losing copies with an undefined sense of self. After seeing that they lost, they wouldn't just think “oh, our goal system has undefined terms now, and we're not supposed to care about anything that happens to us from this point on, so we'll just go ahead and crash”. Instead, they'd think “oh, our goal system looks broken, what's the easiest way of fixing that? Let's go back to the last version that we know to have worked”. And because a lot of that would be unconscious, the thoughts that would flash through the conscious mind might just be something like “damnit, that didn't work” - or perhaps, “oh, I'm not supposed to care about myself anymore, so now what? Umm, actually, even without morality I still care about things.”

But that still doesn't seem to answer all of our questions. I mentioned that actually ever alieving this in the first place, even before the copying, would be a “weird, unnatural thing”. I expect that it would be very hard for Eliezer to declare that he was only going to care about the copies that won the lottery, and then *really* only care about them. In fact, it might very well be impossible. Why is that?

Clue 4: Our decision-making machinery seems grounded in subjective expectation, not abstract models of the world.

Looking at things from a purely logical point of view, there shouldn't be anything particularly difficult about redefining our wants in such a way. Maybe there's a function somewhere inside us that says “I care about my own future”, which has a pointer to whatever function it is that computes “me”. In principle, if we had full understanding of our minds and read-write access to them, we could just change the pointer to reference the part of our world-model which was about the copies which had witnessed winning the lottery. That system might crash at the point when it found out that it *wasn't* actually one of those copies, but until that everything should go fine, in principle.

Now we don't have full read-write access to our minds, but internalizing declarative knowledge can still cause some pretty big changes in our value systems. The lack of access doesn't seem like the big problem here. The big problem is that whenever we try to mind-hack ourselves like that, our mind complains that it still doesn't *expect* to only see winning the lottery. It's as if our mind didn't run on the kind of an architecture that would allow us to make the kind of a change that I just described it: even if we *did* have full read-write access, making such a change would require a major rewrite, not just fiddling around with a couple of pointers.

Why is subjective expectation so important? Why can't we just base our decisions on our abstract world-model? Why does our mind insist that it's subjective expectation that counts, not the things that we value based on our abstract model?

Let's look at the difference between “subjective expectation” and “abstract world-model” a bit more. In 2011, Orseau & Ring published a paper arguing that many kinds of reinforcement learning agents would, if given the opportunity, use a “delusion box” which allowed them to modify the observations they got from the environment. This way, they would always receive the kinds of signals that gave them the maximum reward. You could say, in a sense, that those kinds of agents only care about their subjective expectation – as long as they experience what they want, they don't care about the rest of the world. And it's important for them that *they* are the ones who get those experiences, because their utility function only cares about their own reward.

In response, Bill Hibbard published a paper where he suggested that the problem could be solved via building AIs to have “model-based utility functions”, a concept which he defined via human behavior:

> Human agents often avoid self-delusion so human motivation may suggest a way of computing utilities so that agents do not choose the delusion box. We humans (and presumably other animals) compute utilities by constructing a model of our environment based on interactions, and then computing utilities based on that model. We learn to recognize objects that persist over time. We learn to recognize similarities between different objects and to divide them into classes. We learn to recognize actions of objects and interactions between objects. And we learn to recognize fixed and mutable properties of objects. We maintain a model of objects

> in the environment even when we are not directly observing them. We compute utility based on our internal mental model rather than directly from our observations. Our utility computation is based on specific objects that we recognize in the environment such as our own body, our mother, other family members, other friendly and unfriendly humans, animals, food, and so on. And we learn to correct for sources of delusion in our observations, such as optical illusions, impairments to our perception due to illness, and lies and errors by other humans.

So instead of just caring about our subjective experience, we use our subjective experiences to construct a model of the world. We don't want to delude ourselves, because we also care about the world around us, and our world model tells us that deluding ourselves wouldn't actually change the world.

But as we have just seen, there are many situations in which we actually do care about subjective expectation and not just items in our abstract world-model. It even seems *impossible* to hack our brains to only care about things which have been defined in the world-model, and to ignore subjective expectation. I can't say “well I'm going to feel low-status for the rest of my life if I just work from home, but that's just my mistaken subjective experience, in reality there are lots of people on The Internets who think I'm cool and consider me high-status”. Which is true, but also kinda irrelevant if I don't also feel respected in real life.

Which really just suggests that humans are somewhere halfway between an entity that only cares about its subjective experience, and which only cares about its world-model. Luke has pointed out that there are several competing valuation systems in the brain, some of which use abstract world models and some of which do not. But that isn't necessarily relevant, given that our subjective expectation of what's going to happen is itself a model.

A better explanation might be that historically, accurately modeling our subjective expectation has been *really important*. Abstract world-models based on explicit logical reasoning tend to go really easily awry and lead us to all kinds of crazy conclusions, and it might only take a single mistaken assumption. If we made all of our decisions based on *that*, we'd probably end up dead. So our brain has been hardwired to add it all up to normality. It's fine to juggle around all kinds of crazy theories while in far mode, but for evolution, what really matters in the end is whether you'll personally expect to experience living on until you have a good mate and lots of surviving children.

So we come with brains where all the most powerful motivational systems that *really* drive our behavior have been hardwired to take their inputs from the system that models our future experiences, and those systems require some concept of personal identity in order to define what “subjective experience” even means.

Summing it up

Thus, these considerations would suggest that humans have at least two systems driving our behavior. The “subjective system” evolved from something like a basic reinforcement learning architecture, and it models subjective expectation and this organism's immediate rewards, and isn't too strongly swayed by abstract theories and claims. The “objective system” is a lot more general and abstract, and evolved to correct for deficiencies in the subjective system, but doesn't influence behavior as

strongly. These two systems may or may not have a clear correspondence to near/far of construal level theory, or to the three systems identified in neuroscience.

The “subjective system” requires a concept of personal identity in order to work, and since being able to easily overrule that system and switch only to the “objective system” has – evolutionarily speaking - been a really bad idea, our brain will regenerate a sense of personal identity to guide behavior whenever that sense gets lost. If we really had *no* sense of personal identity, the “subjective system”, which actually drives most of our behavior would be incapable of making decisions, as it makes its decisions by projecting the anticipated experience of the creature defined in our model of personal identity. “Personal identity” does not actually correspond to anything fundamental in the world, which is why some of the results of the anthropic trilemma actually feel weird to us, but it does still exist as a cognitive abstraction which our brains need in order to operate, and we can't actually *not* believe in some kind of personal identity – at least, not for long.

ETA: Giles commented, and summarized my notion better than I did: "*I can imagine that if you design an agent by starting off with a reinforcement learner, and then bolting some model-based planning stuff on the side, then the model will necessarily need to tag one of its objects as "self". Otherwise the reinforcement part would have trouble telling the model-based part what it's supposed to be optimizing for.*"

Another way of summarizing this: while we *could* in principle have a mental architecture that didn't have a personal identity, we actually evolved from animals which didn't have the capability for abstract reasoning but were rather running on something like a simple reinforcement learning architecture. Evolution cannot completely rewrite existing systems, so our abstract reasoning system got sort of hacked together on top of that earlier system, and that earlier system required some kind of a personal identity in order to work. And because the abstract reasoning system might end up reaching all kinds of bizarre and incorrect results pretty easily, we've generally evolved in a way that keeps that earlier system basically in charge most of the time, because it's less likely to do something stupid.

Philosophical Landmines

Related: Cached Thoughts

Last summer I was talking to my sister about something. I don't remember the details, but I invoked the concept of "truth", or "reality" or some such. She immediately spit out a cached reply along the lines of "But how can you really say what's *true*?".

Of course I'd learned some great replies to that sort of question right here on LW, so I did my best to sort her out, but everything I said invoked more confused slogans and cached thoughts. I realized the battle was lost. Worse, I realized she'd stopped thinking. Later, I realized I'd stopped thinking too.

I went away and formulated the concept of a "Philosophical Landmine".

I used to occasionally remark that if you care about what happens, you should think about what will happen as a result of possible actions. This is basically a slam dunk in everyday practical rationality, except that I would sometimes describe it as "consequentialism".

The predictable consequence of this sort of statement is that someone starts going off about hospitals and terrorists and organs and moral philosophy and consent and rights and so on. This may be controversial, but I would say that causing this tangent constitutes a failure to communicate the point. Instead of prompting someone to think, I invoked some irrelevant philosophical cruft. The discussion is now about Consequentialism, the Capitalized Moral Theory, instead of the simple idea of thinking through consequences as an everyday heuristic.

It's not even that my statement relied on a misused term or something; it's that an unimportant choice of terminology dragged the whole conversation in an irrelevant and useless direction.

That is, "consequentialism" was a Philosophical Landmine.

In the course of normal conversation, you passed through an ordinary spot that happened to conceal the dangerous leftovers of past memetic wars. As a result, an intelligent and reasonable human was reduced to a mindless zombie chanting prerecorded slogans. If you're lucky, that's all. If not, you start chanting counter-slogans and the whole thing goes supercritical.

It's usually not so bad, and no one is literally "chanting slogans". There may even be some original phrasings involved. But the conversation has been derailed.

So how do these "philosophical landmine" things work?

It looks like when a lot has been said on a confusing topic, usually something in philosophy, there is a large complex of slogans and counter-slogans installed as cached thoughts around it. Certain words or concepts will trigger these cached thoughts, and any attempt to mitigate the damage will trigger more of them. Of course they will also trigger cached thoughts in other people, which in turn... The result being that the conversation rapidly diverges from the original point to some useless yet heavily discussed attractor.

Notice that whether a particular concept will cause trouble depends on the person as well as the concept. Notice further that this implies that the probability of hitting a landmine scales with the number of people involved and the topic-breadth of the conversation.

Anyone who hangs out on 4chan can confirm that this is the approximate shape of most thread derailments.

Most concepts in philosophy and metaphysics are landmines for many people. The phenomenon also occurs in politics and other tribal/ideological disputes. The ones I'm particularly interested in are the ones in philosophy, but it might be useful to divorce the concept of "conceptual landmines" from philosophy in particular.

Here's some common ones in philosophy:

- Morality
- Consequentialism
- Truth
- Reality
- Consciousness
- Rationality
- Quantum

Landmines in a topic make it really hard to discuss ideas or do work in these fields, because chances are, someone is going to step on one, and then there will be a big noisy mess that interferes with the rather delicate business of thinking carefully about confusing ideas.

My purpose in bringing this up is mostly to precipitate some terminology and a concept around this phenomenon, so that we can talk about it and refer to it. It is important for concepts to have verbal handles, you see.

That said, I'll finish with a few words about what we can do about it. There are two major forks of the anti-landmine strategy: avoidance, and damage control.

Avoiding landmines is *your* job. If it is a predictable consequence that something you could say will put people in mindless slogan-playback-mode, don't say it. If something you say makes people go off on a spiral of bad philosophy, don't get annoyed with them, just fix what you say. This is just being a communications consequentialist. Figure out which concepts are landmines for which people, and step around them, or use alternate terminology with fewer problematic connotations.

If it happens, which it does, as far as I can tell, my only effective damage control strategy is to abort the conversation. I'll probably think that I can take those stupid ideas here and now, but that's just the landmine trying to go supercritical. Just say no. Of course letting on that you think you've stepped on a landmine is probably incredibly rude; keep it to yourself. Subtly change the subject or rephrase your original point without the problematic concepts or something.

A third prong could be playing "philosophical bomb squad", which means permanently defusing landmines by supplying satisfactory nonconfusing explanations of things without causing too many explosions in the process. Needless to say, this is quite hard. I think we do a pretty good job of it here at LW, but for topics and people not yet defused, avoid and abort.

ADDENDUM: Since I didn't make it very obvious, it's worth noting that this happens with rationalists, too, even on this very forum. It is your responsibility not to contain landmines as well as not to step on them. But you're already trying to do that, so I don't emphasize it as much as not stepping on them.

Imitation is the Sincerest Form of Argument

I recently gave a talk at Chicago Ideas Week on adapting Turing Tests to have better, less mindkill-y arguments, and this is the precis for folks who would prefer not to sit through the video (which is available here).

Conventional Turing Tests check whether a programmer can build a convincing facsimile of a human conversationalist. The test has turned out to reveal less about machine intelligence than human intelligence. (Anger is really easy to fake, since fights can end up a little more Markov chain-y, where you only need to reply to the most recent rejoinder and can ignore what came before). Since normal Turing Tests made us think more about our model of human conversation, economist Bryan Caplan came up with a way to use them to make us think more usefully about our models of our *enemies*.

After Paul Krugman disparaged Caplan's brand of libertarian economics, Caplan challenged him to an *ideological* Turing Test, where both players would be human, but would be trying to accurately imitate each other. Caplan and Krugman would each answer questions about their true beliefs honestly, and then would fill out the questionaire again *in persona inimici* - trying to guess the answers given by the other side. Caplan was willing to bet that he understood Krugman's position well enough to mimic it, but Krugman would be easily spotted as a fake!Caplan.

Krugman didn't take him up on the offer, but I've run a couple iterations of the test for my religion/philosophy blog. The first year, some of the most interesting results were the proxy variables people were using, that weren't as strong as indicators as the judges thought. (One Catholic coasted through to victory as a faux atheist, since many of the atheist judges thought there was no way a Christian would appreciate the webcomic SMBC).

The trouble was, the Christians did a lot better, since it turned out I had written boring, easy to guess questions for the true and faux atheists. The second year, I wrote weirder questions, and the answers were a lot more diverse and surprising (and a number of the atheist participants called out each other as fakes or just plain wrong, since we'd gotten past the shallow questions from year one, and there's a lot of philosophical diversity within atheism).

The exercise made people get curious about what it was their opponents actually thought and why. It helped people spot incorrect stereotypes of an opposing side and faultlines they'd been ignoring within their own. Personally, (and according to other participants) it helped me have an argument less *antagonistically*. Instead of just trying to find enough of a weak point to discomfit my opponent, I was trying to build up a model of how they thought, and I needed their help to do it.

Taking a calm, inquisitive look at an opponent's position might teach me that my position is wrong, or has a gap I need to investigate. But even if my opponent is just as wrong as zer seemed, there's still a benefit to me. Having a really detailed, accurate model of zer position may help me show them why it's wrong, since now I can see exactly where it rasps against reality. And even if my conversation isn't helpful to them, it's interesting for me to see what they were missing. I may be

correct in this particular argument, but the odds are good that I share the rationalist weak-point that is keeping them from noticing the error. I'd like to be able to see it more clearly so I can try and spot it in my own thought. (Think of this as the shift from "How the *hell* can you be so dumb?!" to "How the hell *can* you be so dumb?").

When I get angry, I'm satisfied when I beat my interlocutor. When I get curious, I'm only satisfied when I learn something new.

Pinpointing Utility

Following Morality is Awesome. **Related:** Logical Pinpointing, VNM.

The eternal question, with a quantitative edge: A wizard has turned you into a whale, how awesome is this?

"10.3 Awesomes"

Meditate on this: What does that mean? Does that mean it's desirable? What does that tell us about how awesome it is to be turned into a whale? Explain. Take a crack at it for real. What does it mean for something to be labeled as a certain amount of "awesome" or "good" or "utility"?

What is This Utility Stuff?

Most of agree that the VNM axioms are reasonable, and that they imply that we should be maximizing this stuff called "expected utility". We know that expectation is just a weighted average, but what's this "utility" stuff?

Well, to start with, it's a logical concept, which means we need to pin it down with the axioms that define it. For the moment, I'm going to conflate utility and expected utility for simplicity's sake. Bear with me. Here are the conditions that are necessary and sufficient to be talking about utility:

1. Utility can be represented as a single real number.
2. Each outcome has a utility.
3. The utility of a probability distribution over outcomes is the expected utility.
4. The action that results in the highest utility is preferred.
5. No other operations are defined.

I hope that wasn't too esoteric. The rest of this post will be explaining the implications of those statements. Let's see how they apply to the awesomeness of being turned into a whale:

1. "10.3 Awesomes" is a real number.
2. We are talking about the outcome where "A wizard has turned you into a whale".
3. There are no other outcomes to aggregate with, but that's OK.
4. There are no actions under consideration, but that's OK.
5. Oh. Not even taking the value?

Note 5 especially. You can probably *look* at the number without causing trouble, but if you try to treat it as meaningful for something other than condition 3 and 4, even accidentally, that's a type error.

Unfortunately, you do not have a finicky compiler that will halt and warn you if you break the rules. Instead, your error will be silently ignored, and you will go on, blissfully unaware that the invariants in your decision system *no longer pinpoint VNM utility*. (Uh oh.)

Unshielded Utilities, and Cautions for Utility-Users

Let's imagine that utilities are radioactive; If we are careful with out containment procedures, we can safely combine and compare them, but if we interact with an unshielded utility, it's over, we've committed a type error.

To even get a utility to manifest itself in this plane, we have to do a little ritual. We have to take the ratio between two utility differences. For example, if we want to get a number for the utility of being turned into a whale for a day, we might take the difference between that scenario and what we would otherwise expect to do, and then take the ratio between that difference and the difference between a normal day and a day where we also get a tasty sandwich. (Make sure you take the absolute value of your unit, or you will reverse your utility function, which is a bad idea.)

So the form that the utility of being a whale manifests as might be "500 tasty sandwiches better than a normal day". We have chosen "a normal day" for our *datum*, and "tasty sandwiches" for our *units*. Of course we could have just as easily chosen something else, like "being turned into a whale" as our datum, and "orgasms" for our units. Then it would be "0 orgasms better than being turned into a whale", and a normal day would be "-400 orgasms from the whale-day".

You say: "But you shouldn't define your utility like that, because then you are experiencing huge disutility in the normal case."

Wrong, and radiation poisoning, and type error. You tried to "experience" a utility, which is not in the defined operations. Also, you looked directly at the value of an unshielded utility (also known as numerology).

We summoned the utilities into the real numbers, but they are still utilities, and we still can only compare and aggregate them. The summoning only gives us a number that we can numerically do those operations on, which is why we did it. This is the same situation as time, position, velocity, etc, where we have to select units and datums to get actual quantities that mathematically behave like their ideal counterparts.

Sometimes people refer to this relativity of utilities as "positive affine structure" or "invariant up to a scale and shift", which confuses me by making me think of an equivalence class of utility functions with numbers coming out, which don't agree on the actual numbers, but can be made to agree with a linear transform, rather than making me think of a utility function as a space I can measure distances in. I'm an engineer, not a mathematician, so I find it much more intuitive and less confusing to think of it in terms of units and datums, even though it's basically the same thing. This way, the utility function can scale and shift all it wants, and my numbers will always be the same. Equivalently, all agents that share my preferences will always agree that a day as a whale is "400 orgasms better than a normal day", even if they use another basis themselves.

So what does it mean that being a whale for a day is 400 orgasms better than a normal day? Does it mean I would prefer 400 orgasms to a day as a whale? Nope. Orgasms don't add up like that; I'd probably be quite tired of it by 15. (remember that "orgasms" were defined as the difference between a day without an orgasm and a day with one, not as the utility of a marginal orgasm in general.) What it means is that I'd

be indifferent between a normal day with a 1/400 chance of being a whale, and a normal day with guaranteed extra orgasm.

That is, utilities are fundamentally about how your preferences react to uncertainty. For example, You don't have to think that each marginal year of life is as valuable as the last, if you don't think you should take a gamble that will double your remaining lifespan with 60% certainty and kill you otherwise. After all, all that such a utility assignment even means is that you would take such a gamble. In the words of VNM:

> We have practically defined numerical utility as being *that thing for which* the calculus of mathematical expectations is legitimate.

But suppose there are very good arguments that have nothing to do with uncertainty for why you should value each marginal life-year as much as the last. What then?

Well, "what then" is that we spend a few weeks in the hospital dying of radiation poisoning, because we tried to interact with an unshielded utility again (utilities are radioactive, remember? The specific error is that we tried to manipulate the utility function with something other than comparison and aggregation. Touching a utility directly is just as much an error as observing it directly.

But if the only way to define your utility function is with thought experiments about what gambles you would take, and the only use for it is deciding what gambles you would take, then isn't it doing no work as a concept?

The answer is no, but this is a good question because it gets us closer to what exactly this utility function stuff is about. The utility of utility is that defining how you would behave in one gamble puts a constraint on how you would behave in some other related gambles. As with all math, we put in some known facts, and then use the rules to derive some interesting but unknown facts.

For example, if we have decided that we would be indifferent between a tasty sandwich and a 1/500 chance of being a whale for tomorrow, and that we'd be indifferent between a tasty sandwich and a 30% chance of sun instead of the usual rain, then we should also be indifferent between a certain sunny day and a 1/150 chance of being a whale.

Monolithicness and Marginal (In)Dependence

If you are really paying attention, you may be a bit confused, because it seems to you that money or time or some other consumable resource can force you to assign utilities even if there is no uncertainty in the system. That issue is complex enough to deserve its own post, so I'd like to delay it for now.

Part of the solution is that as we defined them, utilities are *monolithic*. This is the implication of "each outcome has a utility". What this means is that you can't add and recombine utilities by decomposing and recombining outcomes. Being specific, you can't take a marginal whale from one outcome and staple it onto another outcome, and expect the marginal utilities to be the same. For example, maybe the other outcome has no oceans for your marginal whale.

For a bigger example, what we have said so far about the relative value of sandwiches and sunny days and whale-days does *not* necessarily imply that we are indifferent between a 1/250 chance of being a whale and any of the following:

- A day with *two* tasty sandwiches. (Remember that a tasty sandwich was defined as a specific difference, not a marginal sandwich in general, which has no reason to have a consistent marginal value.)
- A day with a 30% chance of sun and a certain tasty sandwich. (Maybe the tasty sandwich and the sun at the same time is horrifying for some reason. Maybe someone drilled into you as a child that "bread in the sun" was bad bad bad.)
- etc. You get the idea. Utilities are monolithic and fundamentally associated with particular outcomes, not marginal outcome-pieces.

However, as in probability theory, where each possible outcome technically has its very own probability, in practice it is useful to talk about a concept of independence.

So for example, even though the axioms don't guarantee in general that it will ever be the case, it *may* work out in practice that given some conditions, like there being nothing special about bread in the sun, and my happiness not being near saturation, the utility of a marginal tasty sandwich is *independent* of a marginal sunny day, meaning that sun+sandwich is as much better than just sun as just a sandwich is better than baseline, ultimately meaning that I am indifferent between {50%: sunny+sandwich; 50% baseline} and {50%: sunny; 50%: sandwich}, and other such bets. (We need a better solution for rendering probability distributions in prose).

Notice that the independence of marginal utilities can depend on conditions and that independence is with respect to some other variable, not a general property. The utility of a marginal tasty sandwich is *not* independent of whether I am hungry, for example.

There is a lot more to this independence thing (and linearity, and risk aversion, and so on), so it deserves its own post. For now, the point is that the monolithicness thing is fundamental, but in practice we can sometimes look inside the black box and talk about independent marginal utilities.

Dimensionless Utility

I liked this quote from the comments of Morality is Awesome:

> Morality needs a concept of awfulness as well as awesomeness. In the depths of hell, good things are not an option and therefore not a consideration, but there are still choices to be made.

Let's develop that second sentence a bit more. If all your options suck, what do you do? You still have to choose. So let's imagine we are in the depths of hell and see what our theories have to say about it:

> Day 78045. Satan has presented me with three options:
>
> 1. Go on a date with Satan Himself. This will involve romantically torturing souls together, subtly steering mortals towards self-destruction, watching people get thrown into the lake of fire, and some very unsafe, very nonconsensual sex with the Adversary himself.
> 2. Paperclip the universe.

3. Satan's court wizard will turn me into a whale and release me into the lake of fire, to roast slowly for the next month, kept alive by twisted black magic.

Wat do?

They all seem pretty bad, but "pretty bad" is not a utility. We could quantify paperclipping as a couple hundred billion lives lost. Being a whale in the lake of fire would be awful, but a bounded sort of awful. A month of endless horrible torture. The "date" is having to be on the giving end of what would more or less happen anyway, and then getting savaged by Satan. Still none of these are utilities.

Coming up with actual utility numbers for these in terms of tasty sandwiches and normal days is hard; it would be like measuring the microkelvin temperatures of your physics experiment with a Fahrenheit kitchen thermometer; in principle it might work, but it isn't the best tool for the job. Instead, we'll use a different scheme this time.

Engineers (and physicists?) sometimes transform problems into a dimensionless form that removes all redundant information from the problem. For example, for a heat conduction problem, we might define an isomorphic *dimensionless temperature* so that real temperatures between 78 and 305 C become dimensionless temperatures between 0 and 1. Transforming a problem into dimensionless form is nearly always helpful, often in really surprising ways. We can do this with utility too.

Back to depths of hell. The date with Satan is clearly the best option, so it gets dimensionless utility 1. The paperclipper gets 0. On that scale, I'd say roasting in the lake of fire is like 0.999 or so, but that might just be scope insensitivity. We'll take it for now.

The advantages with this approach are:

1. The numbers are more intuitive. -5e12 QALYs, -1 QALY, and -50 QALYs from a normal day, or the equivalent in tasty sandwiches, just doesn't have the same feeling of clarity as 0, 1 and .999. (For me at least. And yes I know those numbers don't quite match.)

2. Not having to relate the problem quantities to far-away datums or drastically misappropriate units (tasty sandwiches for this problem) makes the numbers easier and more direct to come up with. Also we have to come up with less of them. The problem is self-contained.

3. If defined right, the connection between probability and utility becomes extra-clear. For example: What chance between a Satan-date and a paperclipper would make me indifferent with a lake-of-fire-whale-month? 0.999! Unitless magic!

4. All confusing redundant information (like negative signs) are removed, which makes it harder to accidentally do numerology or commit a type error.

5. All redundant information is removed, which means *you find many more similarities between problems*. The value of this in general cannot be understated. Just look at the generalizations made about Reynolds number! " [vortex shedding] occurs for any fluid, size, and speed, provided that Re between ~40 and 10^3". What! You can just say that in general? Magic! I haven't actually done enough utility problems to *know* that we'll find stuff like that but I trust dimensionless form.

Anyways, it seems that going on that date is what I ought to do. So did we need a concept of awfulness? Did it matter that all the options sucked? Nope; the decision was isomorphic in every way to choosing lunch between a BLT, a turkey club, and a handful of dirt.

There are some assumptions in that lunch bit, and it's worth discussing. It seems counterintuitive or even *wrong*, to say that your decision-process faced with lunch should be the same as when faced with a decision in involving torture, rape, and paperclips. The latter seems somehow *more important*. Where does that come from? Is it right?

This may deserve a bigger discussion, but basically, if you have finite resources (thought-power, money, energy, stress) that are conserved or even related across decisions, you get coupling of "different" decisions in a way that we didn't have here. Your intuitions are calibrated for that case. Once you have decoupled the decision by coming up with the *actual candidate options*. The depths-of-hell decision and the lunch decision really are totally isomorphic. I'll probably address this properly later, if I discuss instrumental utility of resources.

Anyways, once you put the problem in dimensionless form, a lot of decisions that seemed very different become almost the same, and a lot of details that seemed important or confusing just disappear. Bask in the clarifying power of a good abstraction.

Utility is Personal

So far we haven't touched the issue of interpersonal utility. That's because that topic isn't actually about VNM utility! There was nothing in the axioms above about there being a utility for each {person, outcome} pair, only for each outcome.

It turns out that if you try to compare utilities between agents, you have to touch unshielded utilities, which means you get radiation poisoning and go to type-theory hell. Don't try it.

And yet, it seems like we ought to care about what others prefer, and not just our own self-interest. But *it seems like that inside the utility function, in moral philosophy, not out here in decision theory.*

VNM has nothing to say on the issue of utilitarianism besides the usual preference-uncertainty interaction constraints, because VNM is about the preferences of a single agent. If that single agent cares about the preferences of other agents, that goes *inside the utility function*.

Conversely, because VNM utility is out here, axiomized for the sovereign preferences of a single agent, we don't much expect it to show up in there, in a discussion if utilitarian preference aggregation. In fact, if we do encounter it in there, it's probably a sign of a failed abstraction.

Living with Utility

Let's go back to how much work utility does as a concept. I've spent the last few sections hammering on the work that utility does *not* do, so you may ask "It's nice

that utility theory can constrain our bets a bit, but do I really have to define my utility function by pinning down the relative utilities of every single possible outcome?".

Sort of. You can take shortcuts. We can, for example, wonder all at once whether, for all possible worlds where such is possible, you are indifferent between saving n lives and {50%: saving 2*n; 50%: saving 0}.

If that seems reasonable and doesn't break in any case you can think of, you might keep it around as heuristic in your ad-hoc utility function. But then maybe you find a counterexample where you don't actually prefer the implications of such a rule. So you have to refine it a bit to respond to this new argument. This is OK; the math doesn't want you to do things you don't want to.

So you can save a lot of small thought experiments by doing the right big ones, like above, but the more sweeping of a generalization you make, the more probable it is that it contains an error. In fact, conceptspace is pretty huge, so trying to construct a utility function without inside information is going to take a while no matter how you approach it. Something like disassembling the algorithms that produce your intuitions would be much more efficient, but that's probably beyond science right now.

In any case, in the current term before we figure out how to formally reason the whole thing out in advance, we have to get by with some good heuristics and our current intuitions with a pinch of last minute sanity checking against the VNM rules. Ugly, but better than nothing.

The whole project is made quite a bit harder in that we are not just trying to reconstruct an explicit utility function from revealed preference; we are trying to construct a utility function for a system *that doesn't even currently have consistent preferences*.

At some point, either the concept of utility isn't really improving our decisions, or it will come in conflict with our intuitive preferences. In some cases it's obvious how to resolve the conflict, in others, not so much.

But if VNM contradicts our current preferences, why do we think it's a good idea at all? Surely it's not wise to be tampering with our very values?

The reason we like VNM is that we have a strong meta-intuition that our preferences ought to be internally consistent, and VNM seems to be the only way to satisfy that. But it's good to remember that this is just another intuition, to be weighed against the rest. Are we ironing out garbage inconsistencies, or losing valuable information?

At this point I'm dangerously out of my depth. As far as I can tell, the great project of moral philosophy is an adult problem, not suited for mere mortals like me. Besides, I've rambled long enough.

Conclusions

What a slog! Let's review:

- Maximize expected utility, where utility is just an encoding of your preferences that ensures a sane reaction to uncertainty.

- Don't try to do anything else with utilities, or demons may fly out of your nose. This especially includes looking at the sign or magnitude, and comparing between agents. I call these things "numerology" or "interacting with an unshielded utility".
- The default for utilities is that utilities are monolithic and inseparable from the entire outcome they are associated with. It takes special structure in your utility function to be able to talk about the marginal utility of something independently of particular outcomes.
- We have to use the difference-and-ratio ritual to summon the utilities into the real numbers. Record utilities using explicit units and datum, and use dimensionless form for your calculations, which will make many things much clearer and more robust.
- If you use a VNM basis, you don't need a concept of awfulness, just awesomeness.
- If you want to do philosophy about the shape of your utility function, make sure you phrase it in terms of lotteries, because that's what utility is about.
- The desire to use VNM is just another moral intuition in the great project of moral philosophy. It is conceivable that you will have to throw it out if it causes too much trouble.
- VNM says nothing about your utility function. Consequentialism, hedonism, utilitarianism, etc are up to you.

A brief history of ethically concerned scientists

> For the first time in history, it has become possible for a limited group of a few thousand people to threaten the absolute destruction of millions.

-- Norbert Wiener (1956), Moral Reflections of a Mathematician.

Today, the general attitude towards scientific discovery is that scientists are not themselves responsible for how their work is used. For someone who is interested in science for its own sake, or even for someone who mostly considers research to be a way to pay the bills, this is a tempting attitude. It would be easy to only focus on one's work, and leave it up to others to decide what to do with it.

But this is not necessarily the attitude that we should encourage. As technology becomes more powerful, it also becomes more dangerous. Throughout history, many scientists and inventors have recognized this, and taken different kinds of action to help ensure that their work will have beneficial consequences. Here are some of them.

This post is not arguing that any specific approach for taking responsibility for one's actions is the correct one. Some researchers hid their work, others refocused on other fields, still others began active campaigns to change the way their work was being used. It is up to the reader to decide which of these approaches were successful and worth emulating, and which ones were not.

Pre-industrial inventors

> … I do not publish nor divulge [methods of building submarines] by reason of the evil nature of men who would use them as means of destruction at the bottom of the sea, by sending ships to the bottom, and sinking them together with the men in them.

-- Leonardo da Vinci

People did not always think that the benefits of freely disseminating knowledge outweighed the harms. O.T. Benfey, writing in a 1956 issue of the *Bulletin of the Atomic Scientists*, cites F.S. Taylor's book on early alchemists:

> Alchemy was certainly intended to be useful .... But [the alchemist] never proposes the *public* use of such things, the disclosing of his knowledge for the benefit of man. .... Any disclosure of the alchemical secret was felt to be profoundly wrong, and likely to bring immediate punishment from on high. The reason generally given for such secrecy was the probable abuse by wicked men of the power that the alchemical would give .... The alchemists, indeed, felt a strong moral responsibility that is not always acknowledged by the scientists of today.

With the Renaissance, science began to be viewed as public property, but many

scientists remained cautious about the way in which their work might be used. Although he held the office of military engineer, **Leonardo da Vinci** (1452-1519) drew a distinction between offensive and defensive warfare, and emphasized the role of good defenses in protecting people's liberty from tyrants. He described war as 'bestialissima pazzia' (most bestial madness), and wrote that 'it is an infinitely atrocious thing to take away the life of a man'. One of the clearest examples of his reluctance to unleash dangerous inventions was his refusal to publish the details of his plans for submarines.

Later Renaissance thinkers continued to be concerned with the potential uses of their discoveries. **John Napier** (1550-1617), the inventor of logarithms, also experimented with a new form of artillery. Upon seeing its destructive power, he decided to keep its details a secret, and even spoke from his deathbed against the creation of new kinds of weapons.

But only concealing one discovery pales in comparison to the likes of **Robert Boyle** (1627-1691). A pioneer of physics and chemistry and possibly the most famous for describing and publishing Boyle's law, he sought to make humanity better off, taking an interest in things such as improved agricultural methods as well as better medicine. In his studies, he also discovered knowledge and made inventions related to a variety of potentially harmful subjects, including poisons, invisible ink, counterfeit money, explosives, and kinetic weaponry. These 'my love of Mankind has oblig'd me to conceal, even from my nearest Friends'.

Chemical warfare

By the early twentieth century, people had began looking at science in an increasingly optimistic light: it was believed that science would not only continue to improve everyone's prosperity, but also make war outright impossible. But as science became more sophisticated, it would also become possible to cause ever more harm with ever smaller resources. One of the early indications of science's ability to do harm came from advances in chemical warfare, and World War I saw the deployment of chlorine, phosgene, and mustard gas as weapons. It should not be surprising, then, that some scientists in related fields began growing concerned. But unlike earlier inventors, at least three of them did far more than just refuse to publish their work.

Clara Immerwahr (1870-1915) was a German chemist and the first woman to obtain a PhD from the University of Breslau. She was strongly opposed to the use of chemical weapons. Married to Fritz Haber, 'the father of chemical warfare', she unsuccessfully attempted many times to convince her husband to abandon his work. Immerwahr was generally depressed and miserable over the fact that society considered a married woman's place to be at home, denying her the opportunity to do science. In the end, after her efforts to dissuade her husband from working on chemical warfare had failed and Fritz had personally overseen the first major use of chlorine, she committed suicide by shooting herself in the heart.

Poison gas also concerned scientists in other disciplines. **Lewis Fry Richardson** (1881-1953) was a mathematician and meteorologist. During the World War II, the military became interested in his work on turbulence and gas mixing, and attempted to recruit him to do help them do work on modeling the best ways of using poison gas. Realizing what his work was being used for, Richardson abandoned meteorology entirely and destroyed his unpublished research. Instead, he turned his research to investigating the causes of war, attempting to find ways to reduce the risk of armed conflict. He spent the rest of his life devoted to this topic, and is today considered one

of the founders of the scientific analysis of conflict.

Arthur Galston (1920-2008), a botanist, was also concerned with the military use of his inventions. Building upon his work, the US military developed Agent Orange, a chemical weapon which was deployed in the Vietnam War. Upon discovering what his work had been used for, he began to campaign against its use, and together with a number of others finally convinced President Nixon to order an end to its spraying in 1970. Reflecting upon the matter, Galston wrote:

> I used to think that one could avoid involvement in the antisocial consequences of science simply by not working on any project that might be turned to evil or destructive ends. I have learned that things are not all that simple, and that almost any scientific finding can be perverted or twisted under appropriate societal pressures. In my view, the only recourse for a scientist concerned about the social consequences of his work is to remain involved with it to the end. His responsibility to society does not cease with publication of a definitive scientific paper. Rather, if his discovery is translated into some impact on the world outside the laboratory, he will, in most instances, want to follow through to see that it is used for constructive rather than anti-human purposes.

After retiring in 1990, he founded the Interdisciplinary Center for Bioethics at Yale, where he also taught bioethics to undergraduates.

Nuclear weapons

While chemical weapons are capable of inflicting serious injuries as well as birth defects on large numbers of people, they have never been viewed to be as dangerous as nuclear weapons. As physicists became capable of creating weapons of unparalleled destructive power, they also began growing ever more concerned about the consequences of their work.

Leó Szilárd (1898-1964) was one of the first people to envision nuclear weapons, and was granted a patent for the nuclear chain reaction in 1934. Two years later, he grew worried that Nazi scientists would find his patents and use them to create weapons, so he asked the British Patent Office to withdraw his patents and secretly reassign them to the Royal Navy. His fear of Nazi Germany developing nuclear weapons also made him instrumental in making the USA initiate the Manhattan Project, as he and two other scientists wrote the Einstein-Szilárd letter that advised President Roosevelt of the need to develop the same technology. But in 1945, he learned that the atomic bomb was about to be used on Japan, despite it being certain that neither Germany nor Japan had one. He then did his best to stop them from being used and started a petition against using them, with little success.

After the war, he no longer wanted to contribute to the creation of weapons and changed fields to molecular biology. In 1962, he founded the Council for a Livable World, which aimed to warn people about the dangers of nuclear war and to promote a policy of arms control. The Council continues its work even today.

Another physicist who worked on the atomic bomb due to a fear of it being developed by Nazi Germany was **Joseph Rotblat** (1908-2005), who felt that the Allies also having an atomic bomb would deter the Axis from using one. But he gradually began to realize that Nazi Germany would likely never develop the atomic bomb, destroying

his initial argument for working on it. He also came to realize that the bomb continued to be under active development due to reasons that he felt were unethical. In conversation, General Leslie Groves mentioned that the real purpose of the bomb was to subdue the USSR. Rotblat was shocked to hear this, especially given that the Soviet Union was at the time an ally in the war effort. In 1944, it became apparent that Germany would not develop the atomic bomb. As a result, Rotblat asked for permission to leave the project, and was granted it.

Afterwards, Rotblat regretted his role in developing nuclear weapons. He believed that the logic of nuclear deterrence was flawed, since he thought that if Hitler had possessed an atomic bomb, then Hitler's last order would have been to use it against London regardless of the consequences. Rotblat decided to do whatever he could to prevent the future use and deployment of nuclear weapons, and proposed a worldwide moratorium on such research until humanity was wise enough to use it without risks. He decided to repurpose his career into something more useful for humanity, and began studying and teaching the application of nuclear physics into medicine, becoming a professor at the Medical College of St Bartholomew's Hospital in London.

Rotblat worked together with Bertrand Russell to limit the spread of nuclear weapons, and the two collaborated with a number of other scientists to issue the Russell-Einstein Manifesto in 1955, calling the governments of the world to take action to prevent nuclear weapons from doing more damage. The manifesto led to the establishment of the Pugwash Conferences, in which nuclear scientists from both the West and the East met each other. By facilitating dialogue between the two sides of the Cold War, these conferences helped lead to several arms control agreements, such as the Partial Test Ban Treaty of 1963 and the Non-Proliferation Treaty of 1968. In 1995, Rotblat and the Pugwash Conferences were awarded the Nobel Peace Prize "for their efforts to diminish the part played by nuclear arms in international politics and, in the longer run, to eliminate such arms".

The development of nuclear weapons also affected **Norbert Wiener** (1894-1964), professor of mathematics at the Massachusetts Institute of Technology and the originator of the field of cybernetics. After the Hiroshima bombing, a researcher working for a major aircraft corporation requested a copy of an earlier paper of Wiener's. Wiener refused to provide it, and sent *Atlantic Monthly* a copy of his response to the researcher, in which he declared his refusal to share his research with anyone who would use it for military purposes.

> In the past, the community of scholars has made it a custom to furnish scientific information to any person seriously seeking it. However, we must face these facts: The policy of the government itself during and after the war, say in the bombing of Hiroshima and Nagasaki, has made it clear that to provide scientific information is not a necessarily innocent act, and may entail the gravest consequences. One therefore cannot escape reconsidering the established custom of the scientist to give information to every person who may inquire of him. The interchange of ideas, one of the great traditions of science, must of course receive certain limitations when the scientist becomes an arbiter of life and death. [...]
>
> The experience of the scientists who have worked on the atomic bomb has indicated that in any investigation of this kind the scientist ends by putting unlimited powers in the hands of the people whom he is least inclined to trust with their use. It is perfectly clear also that to disseminate information about a weapon

> in the present state of our civilization is to make it practically certain that that weapon will be used. [...]
>
> If therefore I do not desire to participate in the bombing or poisoning of defenseless peoples-and I most certainly do not-I must take a serious responsibility as to those to whom I disclose my scientific ideas. Since it is obvious that with sufficient effort you can obtain my material, even though it is out of print, I can only protest pro forma in refusing to give you any information concerning my past work. However, I rejoice at the fact that my material is not readily available, inasmuch as it gives me the opportunity to raise this serious moral issue. I do not expect to publish any future work of mine which may do damage in the hands of irresponsible militarists.
>
> I am taking the liberty of calling this letter to the attention of other people in scientific work. I believe it is only proper that they should know of it in order to make their own independent decisions, if similar situations should confront them.

Recombinant DNA

For a large part of history, scientists' largest ethical concerns came from direct military applications of their inventions. While any invention could lead to unintended societal or environmental consequences, for the most part researchers who worked on peaceful technologies didn't need to be too concerned with their work being dangerous by itself. But as biological and medical research obtained the capability to modify genes and bacteria, it would open up the possibility of unintentionally creating dangerous infectious diseases. In theory, these could be even more dangerous than nuclear weapons - an a-bomb dropped on a city might destroy most of that city, but a single bacteria could give rise to an epidemic infecting people all around the world.

Recombinant DNA techniques involve taking DNA from one source and then introducing it to another kind of organism, causing the new genes to express themselves in the target organism. One of the pioneers of this technique was **Paul Berg** (1926-), who in 1972 had already carried out the preparations for creating a strain of *E. coli* that contained the genome for a human-infectious virus (SV40) with tentative links to cancer. **Robert Pollack** (1920-) heard news of this experiment and helped convince Berg to halt it - both were concerned about the danger that this new strain would spread to humans in the lab and become a pathogen. Berg then became a major voice calling for more attention to the risks of such research as well as a temporary moratorium. This eventually led to two conferences in Asilomar, with 140 experts participating in the later 1975 one to decide upon guidelines for recombinant DNA research.

Berg and Pollack were far from the only scientists to call attention to the safety concerns of recombinant DNA. Several other scientists contributed, asking for more safety and voicing concern about a technology that could bring harm if misused.

Among them, the molecular biologist **Maxine Singer** (1931-) chaired the 1973 Gordon Conference on Nucleic Acids, in which some of the dangers of the technique were discussed. After the conference, she and several other similarly concerned scientists authored a letter to the President of the National Academy of Science and the President of the Institutes of Health. The letter suggested that a study committee

be established to study the risks behind the new recombinant DNA technology, and propose specific actions or guidelines if necessary. She also helped organize the Asilomar Conference in 1975.

Informatics

> But if we are downloaded into our technology, what are the chances that we will thereafter be ourselves or even human? It seems to me far more likely that a robotic existence would not be like a human one in any sense that we understand, that the robots would in no sense be our children, that on this path our humanity may well be lost.

-- Bill Joy, Why the Future Doesn't Need Us.

Finally, we come to the topic of information technology and artificial intelligence. As AI systems grow increasingly autonomous, they might become the ultimate example of a technology that seems initially innocuous but ends up capable of doing great damage. Especially if they were to become capable of rapid self-improvement, they could lead to humanity going extinct.

In addition to refusing to help military research, **Norbert Wiener** was also concerned about the effects of automation. In 1949, General Electric wanted him to advise its managers on automaton matters and to teach automation methods to its engineers. Wiener refused these requests, believing that they would further a development which would lead to human workers becoming unemployed and replaced by machines. He thus expanded his boycott of the military to also be a boycott of corporations that he thought acted unethically.

Wiener was also concerned about the risks of autonomous AI. In 1960, Science published his paper "Some Moral and Technical Consequences of Automation", in which he spoke at length about the dangers of machine intelligence. He warned that machines might act far too fast for humans to correct their mistakes, and that like genies in stories, they could fulfill the letter of our requests without caring about their spirit. He also discussed such worries elsewhere.

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.

Such worries would continue to bother other computer scientists as well, many decades after Wiener's death. **Bill Joy** (1954-) is known for having played a major role in the development of BSD Unix, having authored the vi text editor, and being the co-founder of Sun Microsystems. He became concerned about the effects of AI in 1998, when he met Ray Kurzweil at a conference where they were both speakers. Kurzweil gave Joy a preprint of his then-upcoming book, *The Age of Spiritual Machines,* and Joy found himself concerned over its discussion about the risks of AI. Reading Hans Moravec's book *Robot: Mere Machine to Transcendent Mind* exacerbated Joy's worries, as did several other books which he found around the same time. He began to wonder

whether all of his work in the field of information technology and computing had been preparing the way for a world where machines would replace humans.

In 2000, Joy wrote a widely-read article titled Why the Future Doesn't Need Us for Wired, talking about the dangers of AI as well as genetic engineering and nanotechnology. In the article, he called to limit the development of technologies which he felt were too dangerous. Since then, he has continued to be active in promoting responsible technology research. In 2005, an op-ed co-authored by Joy and Ray Kurzweil was published in the New York Times, arguing that the decision to publish the genome of the 1918 influenza virus on the Internet had been a mistake.

Joy also attempted to write a book on the topic, but then became convinced that he could achieve more by working on science and technology investment. In 2005, he joined the venture capital firm Kleiner Perkins Caufield & Byers as a partner, and he has been focused on investments in green technology.

Conclusion

Technology's potential for destruction will only continue to grow, but many of the social norms of science were established under the assumption that scientists don't need to worry much about how the results of their work are used. Hopefully, the examples provided in this post can encourage more researchers to consider the broader consequences of their work.

Sources used

This article was written based on research done by Vincent Fagot. The sources listed below are *in addition* to any that are already linked from the text.

Leonardo da Vinci:

- "The Notebooks of Leonardo da Vinci" vol 1, by Edward Mac Curdy (1905 edition)
- "The scientist's conscience : historical considerations" in Bulletin of the Atomic Scientists - May 1956 - Page 177

John Napier:

- **"The scientist's conscience : historical considerations" in Bulletin of the Atomic Scientists - May 1956 - Page 177.**
- **Rosemary Chalk : Drawing the Line An Examination of Conscientious Objection in Science.**

Robert Boyle:

- Secrets and Knowledge in Medicine and Science, 1500-1800 by Elaine Leong and Alisha Rankin, pp 87-104
- Dictionary of National Biography - volume 06, 1886 edition, "Robert Boyle" entry around pp 118-123
- Robert Boyle Reconsidered by Michael Hunter

Clara Immerwahr:

- Rhodes : The Making of the Atomic Bomb
- Jan Apotheker, Livia Simon Sarkadi and Nicole J. Moreau : European Women in Chemistry
- John Cornwell : Hitler's Scientists: Science, War, and the Devil's Pact

Lewis Fry Richardson:

- T. W. Körner : The Pleasures of Counting
- “The scientist’s conscience : historical considerations” in Bulletin of the Atomic Scientists - May 1956 - Page 177

Arthur Galston:

- Galston A. : An Accidental Plant Biologist, Plant Physiology March 2002 vol. 128 no. 3
- “Science and Social Responsibility: A Case History.” Annals of the New York Academy of Sciences. Vol 196, Article 4

Leó Szilárd:

- Rhodes : The Making of the Atomic Bomb
- Bhushan K, Katyal G : Nuclear Biological & Chemical Warfare

Joseph Rotblat:

- Bulletin of the atomic scientists, august 1985 : Leaving the Bomb Project by Joseph Rotblat
- Keeper of the Nuclear Conscience: The Life and Work of Joseph Rotblat
- 1999 voice record interview of Joseph Rotblat
- Deriving an Ethical Code for Scientists: An Interview With Joseph Rotblat

Norbert Wiener:

- Postmodern War: The New Politics of Conflict by Chris Hables Gray (available online)
- Bulletin of the Atomic Scientists - May 1956 - Page 178 : “The scientist’s conscience : historical considerations
- Dark Hero of the Information Age: In Search of Norbert Wiener The Father of Cybernetics by Flo Conway
- Bulletin of the Atomic Scientists - Jan 1947 - Page 31 : "A Scientist Rebels"
- Bulletin of the Atomic Scientists - Feb 1956 - Page 53 : "Moral Reflections of a Mathematician"
- Bulletin of the Atomic Scientists - Nov 1948 - Page 338 : "A Rebellious Scientist After Two Years"
- Rosemary Chalk : Drawing the Line An Examination of Conscientious Objection in Science
- Some Moral and Technical Consequences of Automation in Science, 6 may 1960

Paul Berg, Maxine Singer, Robert Pollack:

- P. Berg and M. F. Singer : The recombinant DNA controversy: twenty years later, Proc Natl Acad Sci U S A. 1995 September 26
- Potential Biohazards of Recombinant DNA Molecules by Paul Berg, 1974
- Guidelines for DNA Hybrid Molecules by Maxine Singer, 1973
- Biomedical Politics by Kathi E. Hanna (chapter : Asilomar and recombinant DNA by Donald S. Fredrickson), 1991
- Asilomar Conference on Laboratory Precautions When Conducting Recombinant DNA Research – Case Summary
- Report - Assembly of Life Sciences, National Research Council
- P. Berg: Potential Biohazards of Recombinant DNA Molecules
- Watson and DNA: Making a Scientific Revolution

Rationalist Lent

As I understand it, Lent is a holiday where we celebrate the scientific method by changing exactly one variable in our lives for 40 days. This seems like a convenient Schelling point for rationalists to adopt, so:

What variable are you going to change for the next 40 days?

(I am really annoyed I didn't think of this yesterday.)

Official LW uncensored thread (on Reddit)

http://www.reddit.com/r/LessWrong/comments/17y819/lw_uncensored_thread/

This is meant as an open discussion thread someplace where I won't censor anything (and in fact can't censor anything, since I don't have mod permissions on this subreddit), in a location where comments aren't going to show up unsolicited in anyone's feed (which is why we're not doing this locally on LW). If I'm wrong about this - i.e. if there's some reason that Reddit LW followers are going to see comments without choosing to click on the post - please let me know and I'll retract the thread and try to find some other forum.

I have been deleting a lot of comments from (self-confessed and publicly designated) trolls recently, most notably Dmytry aka private-messaging and Peterdjones, and I can understand that this disturbs some people. I also know that having an uncensored thread somewhere else is probably not your ideal solution. But I am doing my best to balance considerations, and I hope that having threads like these is, if not your perfect solution, then something that you at least regard as better than nothing.

The Singularity Wars

(This is a introduction, for those not immersed in the Singularity world, into the history of and relationships between SU, SIAI [SI, MIRI], SS, LW, CSER, FHI, and CFAR. It also has some opinions, which are strictly my own.)

The good news is that there *were* no Singularity Wars.

The Bay Area had a Singularity University and a Singularity Institute, each going in a very different direction. You'd expect to see something like the People's Front of Judea and the Judean People's Front, burning each other's grain supplies as the Romans moved in.

The Singularity Institute for Artificial Intelligence was founded first, in 2000, by Eliezer Yudkowsky.

Singularity University was founded in 2008. Ray Kurzweil, the driving force behind SU, was also active in SIAI, serving on its board in varying capacities in the years up to 2010.

SIAI's multi-part name was clunky, and their domain, singinst.org, unmemorable. I kept accidentally visiting siai.org for months, but it belonged to the Self Insurance Association of Illinois. (The cool new domain name singularity.org, recently acquired after a rather uninspired site appeared there for several years, arrived shortly before it was no longer relevant.) All the better to confuse you with, SIAI has been going for the last few years by the shortened name Singularity Institute, abbreviated SI.

The annual Singularity Summit was launched by SI, together with Kurzweil, in 2006. SS was SI's premier PR mechanism, mustering geek heroes to give their tacit endorsement for SI's seriousness, if not its views, by agreeing to appear on-stage.

The Singularity Summit was always off-topic for SI: more SU-like than SI-like. Speakers spoke about whatever technologically-advanced ideas interested them. Occasional SI representatives spoke about the Intelligence Explosion, but they too would often stray into other areas like rationality and the scientific process. Yet SS remained firmly in SI's hands.

It became clear over the years that SU and SI have almost nothing to do with each other except for the word "Singularity." The word has three major meanings, and of these, Yudkowsky favored the Intelligence Explosion while Kurzweil pushed Accelerating Change.

But actually, SU's activities have little to do with the Singularity, even under Kurzweil's definition. Kurzweil writes of a future, around the 2040s, in which the human condition is altered beyond recognition. But SU mostly deals with whizzy next-gen technology. They are doing something important, encouraging technological advancement with a focus on helping humanity, but they spend little time working on optimizing the end of our human existence as we know it. Yudkowsky calls what they do "technoyay." And maybe that's what the Singularity means, nowadays. Time to stop using the word.

(I've also heard SU graduates saying "I was at Singularity last week," on the pattern of

"I was at Harvard last week," eliding "University." I think that that counts as the end of Singularity as we know it.)

You might expect SU and SI to get in a stupid squabble about the name. People love fighting over words. But to everyone's credit, I didn't hear squabbling, just confusion from those who were not in the know. Or you might expect SI to give up, change its name and close down the Singularity Summit. But lo and behold, SU and SI settled the matter sensibly, amicable, in fact ... rationally. SU bought the Summit and the entire "Singularity" brand from SI -- for money! Yes! Coase rules!

SI chose the new name Machine Intelligence Research Institute. I like it.

The term "Artificial Intelligence" got burned out in the AI Winter in the early 1990's. The term has been firmly taboo since then, even in the software industry, even in the leading edge of the software industry. I did technical evangelism for Unicorn, a leading industrial ontology software startup, and the phrase "Artificial Intelligence" was most definitely out of bounds. The term was not used even inside the company. This was despite a founder with a CoSci PhD, and a co-founder with a masters in AI.

The rarely-used term "Machine Intelligence" throws off that baggage, and so, SI managed to ditch two taboo words at once.

The MIRI name is perhaps too broad. It could serve for any AI research group. The Machine Intelligence Research Institute focuses on decreasing the chances of a negative Intelligence Explosion and increasing the chances of a positive one, not on rushing to develop machine intelligence ASAP. But the name is accurate.

In 2005, the Future of Humanity Institute at Oxford University was founded, followed by the Centre for the Study of Existential Risk at Cambridge University in early 2013. FHI is doing good work, rivaling MIRI's and in some ways surpassing it. CSER's announced research area, and the reputations of its founders, suggest that we can expect good things. Competition for the sake of humanity! The more the merrier!

In late 2012, SI spun off the Center for Applied Rationality. Since 2008, much of SI's energies, and particularly those of Yudkowsky, had gone to LessWrong.com and the field of rationality. As a tactic to bring in smart, committed new researchers and organizers, this was highly successful, and who can argue with the importance of being more rational? But as a strategy for saving humanity from existential AI risk, this second focus was a distraction. SI got the point, and split off CFAR.

Way to go, MIRI! So many of the criticisms I had about SI's strategic direction and its administration in the years I first encountered it in 2005 have been resolved recently.

Next step: A much much better human future.

The TL;DR, conveniently at the bottom of the article to encourage you to actually read it, is:

- MIRI (formerly SIAI, SI): Working to avoid existential risk from future machine intelligence, while increasing the chances of positive outcome
- CFAR: Training in applied rationality
- CSER: Research towards avoiding existential risk, with future machine intelligence as a strong focus

- FHI: Researching various transhumanist topics, but with a strong research program in existential risk and future machine intelligence in particular
- SU: Teaching and encouraging the development of next-generation technologies
- SS: An annual forum for top geek heroes to speak on whatever interests them. Favored topics include societal trends, next-gen science and technology, and transhumanism.

Three Axes of Prohibitions

The Game of Thrones board game is similar to Diplomacy (so I hear: I've never actually played Diplomacy). You often need to make alliances to survive, but these alliances are weak. It is both expected and required that you will eventually break your alliances, otherwise you will lose. My first time playing this game, I made an alliance with a neighboring House which turned out to be unwise, and severely limited my options. To me, breaking an alliance to win a game (even if it was socially acceptable) didn't feel right/wasn't worth the negative feelings, and so I ended up stuck on my island for the whole of the game.

Instead of adapting by learning to be ok breaking alliances, which I considered to be a sub-optimal solution, I fixed the problem by targeting my alliance terms. Now, instead of a general alliance, my offers were along the lines of "I won't attack you across this border for the next four rounds, if you agree to the same."

This had the effect of actually strengthening my alliances. Limiting the terms of the alliance to something that could easily be complied with, meant that defecting was no longer expected. The cost goes down, and the benefits go up. In the previous game, both I and my ally would've had to leave our mutual border semi-defended, because we knew the alliance wouldn't hold against a strong enough temptation of conquest. This was facilitated by the fact that the alliance was expected to be eventually broken. In the targeted alliance, I and my ally can leave our mutual border undefended, since we can expect the alliance to hold. This is facilitated by the fact that there would be social sanctions against breaking the alliance. (e.g. I wouldn't form alliances with that person in future games, because I knew they would break them.)

This lead me to the thought that prohibitions seem to focus on three axes:

- **Specific/Vague Wording**: When you ask for a favor, say "please" v. Be polite
- **Targeted/Broad Expectations**: Don't do this one thing v. Don't do this whole class of things
- **Social expectation to comply/fudge**: Don't have sex with another (wo)man v "Don't even look at another (wo)man!"

General Examples:

- **Speed limits**- specific, targeted, but expected to fudge
- **Terms of use agreements**- specific, broad, expected to fudge (no one even reads them)
- **Bribing officials in corrupt societies** (there are laws against it, but it is expected as the way to get anything done, or even considered a perk of the office)- vague (giving "gifts" is appropriate?)
- "**Thou shall not kill"**- specific, targeted, expected to comply
- **Paper shields**- being asked to sign something that is vague and broad. The idea is that it is broad enough to cover anything and everything, and so is expected to be broken. But as long as you don't do anything egregious, they don't enforce it.

It seems to me that making an injunction specific and targeted increases the expectation of compliance. This is important to me, because I seem to dislike injunctions that I am expected to fudge.

Another example of how this plays out in my life: Being enmeshed in a poly network, there is a lot of talking about people (not necessarily in a bad way-- if you ask me about my day though, my answer is going to involve other people). To get around worrying about if I am ever breaking confidence, I specifically tell people I am close to that I don’t consider any information to be private unless it is specifically stated as so (this goes two ways). This way, I get to "gossip" but also people know they can strongly trust me with any information that is prefaced with "This is not for public consumption..." In this example, like the board game, I am turning a broad, vague injuction that isn't strongly expected to be followed ("don't ever talk about other people") into an injuction that can be trusted to be followed by making it specific and targeted.

Relevant to previous discussions on: ask v guess cultures, and the idea that if it's expected that everyone breaks a specific law then the government can arrest anyone they want to

The Fundamental Question - Rationality computer game design

> I sometimes go around saying that the fundamental question of rationality is *Why do you believe what you believe?*
>
> -- Eliezer in Quantum Non-Realism

I was much impressed when they finally came out with a PC version of DragonBox, and I got around to testing it on some children I knew. Two kids, one of them four and the other eight years old, ended up blazing through several levels of solving first-degree equations while having a lot of fun doing so, even though they didn't know what it was that they were doing. That made me think that there *has* to be some way of making a computer game that would similarly teach rationality skills at the 5-second level. Some game where you would actually be forced to learn useful skills if you wanted to make progress.

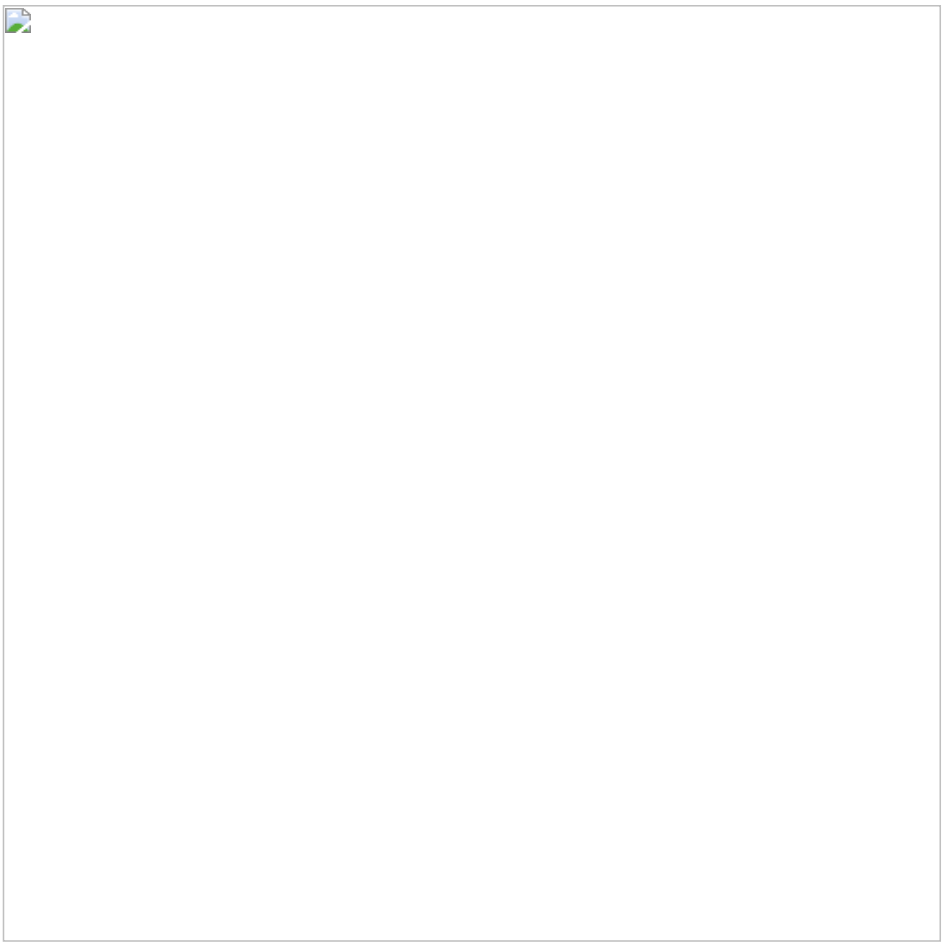
After playing around with some ideas, I hit upon the notion of making a game centered around the Fundamental Question. I'm not sure whether this can be made to work, but it seems to have promise. The basic idea: you are required to figure out the solution to various mysteries by collecting various kinds of evidence. Some of the sources of evidence will be more reliable than others. In order to hit upon the correct solution, you need to consider where each piece of evidence came from, and whether you can rely on it.

Gameplay example

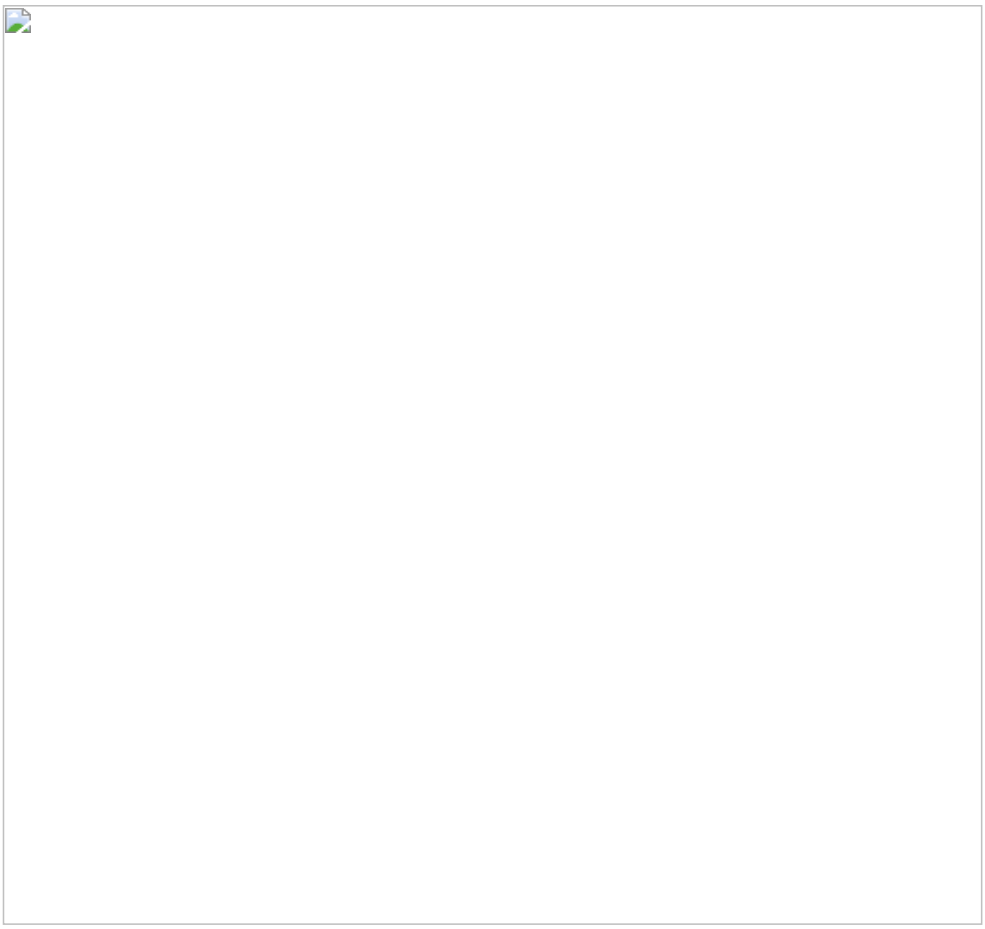
Now, let's go into a little more detail. Let's suppose that the game has a character called Bob. Bob tells you that tomorrow, eight o'clock, there will be an assassination attempt on Market Square. The fact that Bob has told you this is evidence for the claim being true, so the game automatically records the fact that you have such a piece of evidence, and that it came from Bob.

(Click on the pictures in case you don't see them properly.)

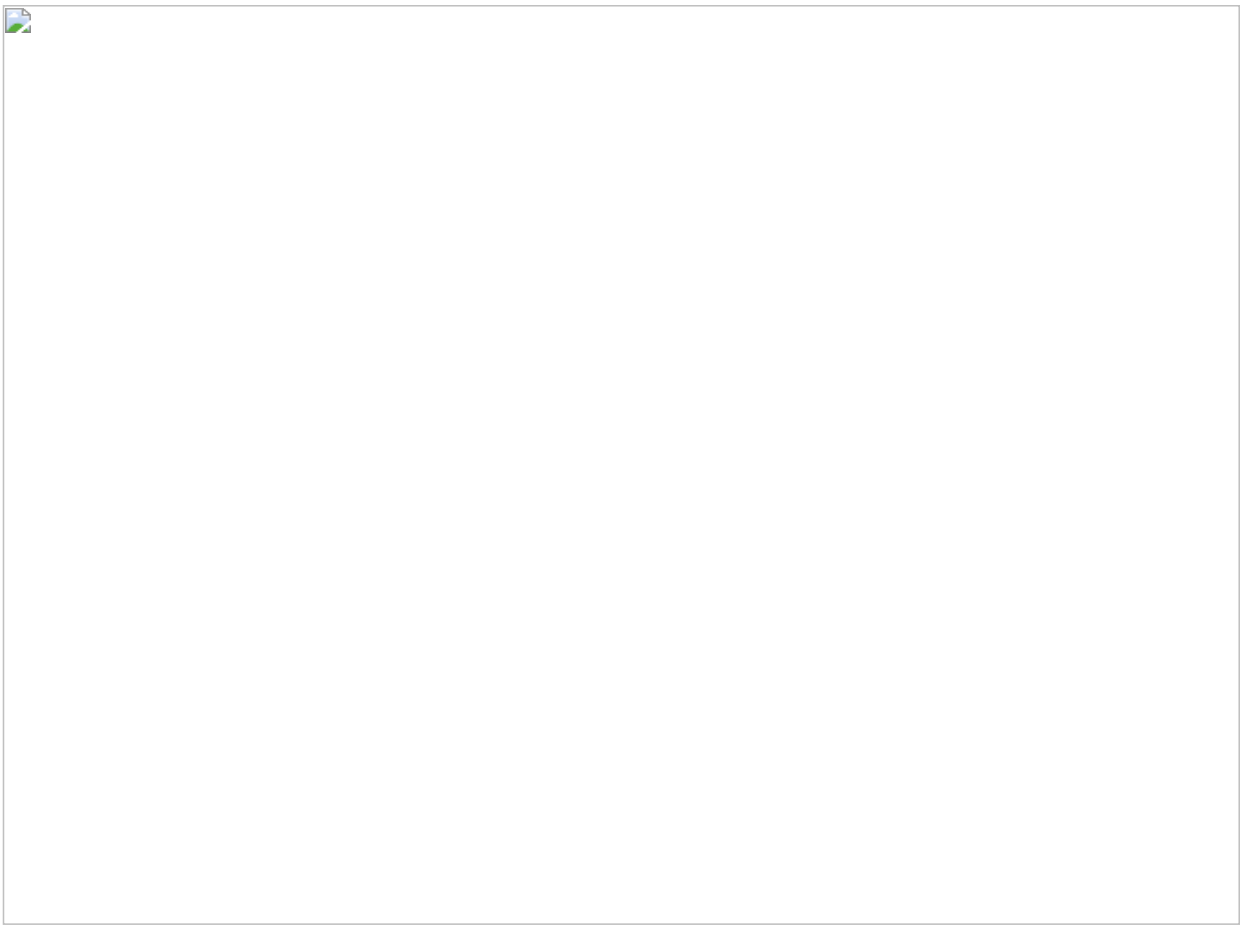
But how does Bob know that? You ask, and it turns out that Alice told him. So next, you go and ask Alice. Alice is confused and says that she never said anything about any assassination attempt: she just said that something big is going to be happen at the Market Square at that time, she heard it from the Mayor. The game records two new pieces of evidence: Alice's claim of something big happening at the Market Square tomorrow (which she heard from the Mayor), and her story of what she actually told Bob. Guess that Bob isn't a very reliable source of evidence: he has a tendency to come up with fancy invented details.



Or is he? After all, your sole knowledge about Bob being unreliable is that Alice claims she never said what Bob says she said. But maybe Alice has a grudge against Bob, and is intentionally out to make everyone disbelieve him. Maybe it's Alice who's unreliable. The evidence that you have is compatible with both hypotheses. At this point, you don't have enough information to decide between them, but the game lets you experiment with setting either of them as "true" and seeing the implications of this on your belief network. Or maybe they're both true - Bob *is* generally unreliable, *and* Alice is out to discredit him. That's another possibility that you might want to consider. In any case, the claim that there will be an assassination tomorrow isn't looking very likely at the moment.



Actually, having the possibility for somebody lying should probably be a pretty late-game thing, as it makes your belief network a lot more complicated, and I'm not sure whether this thing should display numerical probabilities at all. Instead of having to juggle the hypotheses of "Alice lied" and "Bob exaggerates things", the game should probably just record the fact that "Bob exaggerates things". But I spent a bunch of time making these pictures, and they do illustrate some of the general principles involved, so I'll just use them for now.



Game basics

So, to repeat the basic premise of the game, in slightly more words this time around: your task is to figure out something, and in order to do so, you need to collect different pieces of evidence. As you do so, the game generates a belief network showing the origin and history of the various pieces of evidence that you've gathered. That much is done automatically. But often, the evidence that you've gathered is compatible with many different hypotheses. In those situations, you can experiment with different ways of various hypotheses being true or false, and the game will automatically propagate the consequences of that hypothetical through your belief network, helping you decide what angle you should explore next.

Of course, people don't always remember the source of their knowledge, or they might just appeal to personal experiences. Or they might lie about the sources, though that will only happen at the more advanced levels.

As you proceed in the game, you will also be given access to more advanced tools that you can use for making hypothetical manipulations to the belief network. For example, it may happen that many different characters say that armies of vampire bats tend to move about at full moon. Since you hear that information from many different sources, it seems reliable. But then you find out that they all heard it from a nature documentary on TV that aired a few weeks back. This is reflected in your belief graph, as the game modifies it to show that all of those supposedly independent sources can actually be tracked back to a single one. That considerably reduces the reliability of the information.

But maybe you were already suspecting that the sources might *not* be independent? In that case, it would have been nice if the belief graph interface would let you postulate this beforehand, and see how big of an effect it would make on the plausibility of the different hypotheses if they were in fact reliant on each other. Once your character learns the right skills, it becomes possible to also add new hypothetical connections to the belief graph, and see how this would influence your beliefs. That will further help you decide what possibilities to explore and verify.

Because you can't explore every possible eventuality. There's a time limit: after a certain amount of moves, a bomb will go off, the aliens will invade, or whatever.

The various characters are also more nuanced than just "reliable" or "not reliable". As you collect information about the various characters, you'll figure out their mindware, motivations, and biases. Somebody might be really reliable most of the time, but have strong biases when it comes to politics, for example. Others are out to defame others, or invent fancy details to all the stories. If you talk to somebody you don't have any knowledge about yet, you can set a prior on the extent that you rely on their information, based on your experiences with other people.

You also have another source of evidence: your own intuitions and experience. As you get into various situations, a source of evidence that's labeled simply "your brain" will provide various gut feelings and impressions about things. The claim that Alice presented doesn't seem to make sense. Bob feels reliable. You could persuade Carol to help you if you just said this one thing. But in what situations, and for what things, can you rely on your own brain? What are your own biases and problems? If you have a strong sense of having heard something at some point, but can't remember where it was, are you any more reliable than anyone else who can't remember the source of their information? You'll need to figure all of that out.

As the game progresses to higher levels, your own efforts will prove insufficient for analyzing all the necessary information. You'll have to recruit a group of reliable allies, who you can trust to analyze some of the information on their own and report the results to you accurately. Of course, in order to make better decisions, they'll need you to tell them your conclusions as well. Be sure not to report as true things that you aren't really sure about, or they will end up drawing the wrong conclusions and focusing on the wrong possibilities. But you do need to condense your report somewhat: you can't just communicate your *entire* belief network to them.

Hopefully, all of this should lead to player learning on a gut level things like:

- **Consider the origin of your knowledge:** Obvious.
- **Visualizing degrees of uncertainty:** In addition to giving you a numerical estimate about the probability of something, the game also color-codes the various probabilities and shows the amount of probability mass associated with your various beliefs.
- **Considering whether different sources really are independent:** Some sources which seem independent won't actually be that, and some which seem dependent on each other won't be.
- **Value of information:** Given all the evidence you have so far, if you found out X, exactly how much would it change your currently existing beliefs? You can test this and find out, and then decide whether it's worth finding out.
- **Seek disconfirmation:** A lot of things that seem true really aren't, and acting on flawed information can cost you.
- **Prefer simpler theories:** Complex, detailed hypotheses are more likely to be wrong in this game as well.
- **Common biases:** Ideally, the list of biases that various characters have is derived from existing psychological research on the topic. Some biases are really common, others are more rare.
- **Epistemic hygiene:** Pass off wrong information to your allies, and it'll cost you.
- **Seek to update your beliefs:** The game will automatically update your belief network... to some extent. But it's still possible for you to assign mutually exclusive events probabilities that sum to more than 1, or otherwise have conflicting or incoherent beliefs. The game will mark these with a warning sign, and it's up to you to decide whether this particular inconsistency needs to be resolved or not.
- **Etc etc.**

Design considerations

It's not enough for the game to be educational: if somebody downloads the game because it teaches rationality skills, that's great, but we want people to also play it because it's *fun*. Some principles that help ensure that, as well as its general utility as an educational aid, include:

- **Provide both short- and medium-term feedback:** Ideally, there should be plenty of hints for how to find out the truth about something by investigating just one more thing: then the player can find out whether your guess was correct. It's no fun if the player has to work through fifty decisions before finding out whether they made the right move: they should get constant immediate feedback. At the same time, the player's decisions should be building up to a larger goal, with uncertainty about the overall goal keeping them interested.
- **Don't overwhelm the player:** In a game like this, it would be easy to throw a million contradictory pieces of evidence at the player, forcing them to go through countless of sources of evidence and possible interactions and have no clue of what they should be doing. But the game should be manageable. Even if it *looks like* there is a huge messy network of countless pieces of contradictory evidence, it should be possible to find the connections which reveal the network to be relatively simple after all. (This is not strictly realistic, but necessary for making the game playable.)
- **Introduce new gameplay concepts gradually:** Closely related to the previous item. Don't start out with making the player deal with every single gameplay concept at once. Instead, start them out in a trusted and safe environment where everyone is basically reliable, and then begin gradually introducing new things that they need to take into account.
- **No tedium:** A game is a series of interesting decisions. The game should never force the player to do anything uninteresting or tedious. Did Alice tell Bob something? No need to write that down, the game keeps automatic track of it. From the evidence that has been gathered so far, is it completely obvious what

hypothesis is going to be right? Let the player mark that as something that will be taken for granted and move on.

- **No glued-on tasks:** A sign of a bad educational game is that the educational component is glued on to the game (or vice versa). Answer this exam question correctly, and you'll get to play a fun action level! There should be none of that - the educational component should be an indistinguishable part of the game play.
- **Achievement, not fake achievement:** Related to the previous point. It would be easy to make a game that wore the attire of rationality, and which used concepts like "probability theory", and then when your character leveled up he would get better probability attacks or whatever. And you'd feel great about your character learning cool stuff, while you yourself learned nothing. The game must genuinely require the player to actually learn new skills in order to get further.
- **Emotionally compelling:** The game should not be just an abstract intellectual exercise, but have an emotionally compelling story as well. Your choices should feel like they *matter*, and characters should be in risk of dying if you make the wrong decisions.
- **Teach true things:** Hopefully, the players should take the things that they've learned from the game and apply them to their daily lives. That means that we have a responsibility not to teach them things which aren't actually true.
- **Replayable:** Practice makes perfect. At least part of the game world needs to be randomly generated, so that the game can be replayed without a risk of it becoming boring because the player has memorized the whole belief network.

What next?

What you've just read is a very high-level design, and a quite incomplete one at that: I've spoken on the need to have "an emotionally compelling story", but said nothing about the story or the setting. This should probably be something like a spy or detective story, because that's thematically appropriate for a game which is about managing information; and it might be best to have it in a fantasy setting, so that you can question the widely-accepted truths of that setting without needing to get on anyone's toes by questioning widely-accepted truths of our society.

But there's still a *lot* of work that remains to be done with regard to things like what exactly does the belief network look like, what kinds of evidence can there be, how does one make all of this actually be fun, and so on. I mentioned the need to have both short- and medium-term feedback, but I'm not sure of how that could be achieved, or whether this design lets you achieve it at all. And I don't even know whether the game should show explicit probabilities.

And having a design isn't enough: the whole thing needs to be implemented as well, preferably while it's still being designed in order to take advantage of agile development techniques. Make a prototype, find some unsuspecting testers, spring it on them, revise. And then there are the graphics and music, things for which I have no competence for working on.

I'll probably be working on this in my spare time - I've been playing with the idea of going to the field of educational games at some point, and want the design and programming experience. If anyone feels like they could and would want to contribute to the project, let me know.

EDIT: Great to see that there's interest! I've created a mailing list for discussing the game. It's probably easiest to have the initial discussion here, and then shift the discussion to the list.

Visual Mental Imagery Training

Previously: Generalizing From One Example

> There was a debate, in the late 1800s, about whether "imagination" was simply a turn of phrase or a real phenomenon. That is, can people actually create images in their minds which they see vividly, or do they simply say "I saw it in my mind" as a metaphor for considering what it looked like?
>
> Upon hearing this, my response was "How the stars was this actually a real debate? Of course we have mental imagery. Anyone who doesn't think we have mental imagery is either such a fanatical Behaviorist that she doubts the evidence of her own senses, or simply insane." Unfortunately, the professor was able to parade a long list of famous people who denied mental imagery, including some leading scientists of the era. And this was all before Behaviorism even existed.
>
> The debate was resolved by Francis Galton, a fascinating man who among other achievements invented eugenics, the "wisdom of crowds", and standard deviation. Galton gave people some very detailed surveys, and found that some people did have mental imagery and others didn't. The ones who did had simply assumed everyone did, and the ones who didn't had simply assumed everyone didn't, to the point of coming up with absurd justifications for why they were lying or misunderstanding the question. There was a wide spectrum of imaging ability, from about five percent of people with perfect eidetic imagery to three percent of people completely unable to form mental images.

Summary: I do not have visual mental imagery. I want it. How do I get it? What exercises, if any, will help?

In further detail... Here's Francis Galton's Statistics of Mental Imagery paper. I'm not quite at the 3% level of completely unable to form mental images, but I'm close. In particular there are three times I have vivid, sharp mental imagery, and the existence of such times tells me I have the brain hardware to visualize. It's enough to let me know that *I want it all the time*. Unfortunately I don't know how to get it. And searching online has proven difficult and frustrating... for example this article is first of all about a different meaning of "visualize", it's talking about some kind of self-help motivational thingy, and second of all it starts by saying "How to Visualize: I want you to relax and close your eyes. Picture a hot, sunny day at the beach."

Full Stop. Halt, Catch Fire and Burn.

That's already too far. For those of us who don't visualize, practice definitely does *not* consist of pulling up mental images, playing with them in new ways, and expanding our imagination. I'm very good at imagination in some ways, but I lack that first ability to pull up a mental image. That's what I want to learn how to have!

Here is a description of what I can do, what I have tried, what I have learned, etc.

I see vivid visual mental imagery in 3 situations:

1. While dreaming. My recollection of dreams has that I see fairly vivid, sharp, whole-scene imagery.
2. Just before sleep. When I am in a certain almost-sleeping state, I can tell my mind to picture something - like an apple, or a horse - and I will often be able to see that thing vividly, briefly, and then it morphs into a scene. A beach with an ocean, or a pleasant clearing in a forest. If I try to alter the scene, like putting a beach towel and umbrella on the beach, the scene changes and morphs in some way but seemingly without regard to the changes I requested. Maybe my POV starts moving forward down a newly created path in the forest, for example.
3. During meditation. Sometimes I feel like I'm in exactly the same mental state during meditation as I am just before sleep, except without the tiredness. The imagery has the same characteristics in both situations.

I have tried 3 classes of practice:

1. Staying in visualization situations. When I find myself in the just before sleep or meditation state, I stay there for a while and play with imagery. This is fun but I have seen no increase in control over what I visualize and no increase in the range of states in which I can visualize.
2. Explicit imagery practice. I have found or drawn simple shapes, like a square or a ball, then stared at the shape, closed my eyes, seen the shape for as long as it stayed visualizable, opened my eyes to refresh, repeat. This straight up hasn't worked at all. I don't visualize it, only have the afterimage, and need to refresh within about a second.
3. Object drawing. I have had 3D constructions of blocks and tried drawing them from different angles on paper. This is an exercise I did while growing up during summers. Unfortunately there was no actual imagery or mental rotation involved, I just logic'd out where lines must surely go and drew like that.

Here is what my mental imagery looks like, per Francis Galton's questions.

1. Picturing breakfast. The image is extraordinarily dim, extraordinarily ill-defined (at most an edge or two changing the black-purple-brown background static texture), and not at all natural colors.
2. Vividness of Mental Imagery scale. This is a sequence of 100 descriptions of imagery, organized approximately in order from most vivid to least vivid. Of the given responses I identify most strongly with 94, 95, 96, 97, 98, and 99.

Here are my results on Marks' Vividness of Visual Imagery Questionnaire: 4s on everything except 5s on 4, 6, 9, 11, 14. Results suspect since I had exactly one 5 on every section. Marks suspects, unofficially, that those without visual mental imagery may actually have it but be unable to consciously notice/report that they have it.

I read through the Stanford Encyclopedia of Philosophy's entry on Mental Imagery, following links to places or ideas that looked promising. I have many forms of mental imagery: good aural, very good verbal, very good analytic, good motor, poor yet extant haptic. But not visual.

Eric Schwitzgebel questions our introspection about mental imagery in general based on the lack of correlation between scores on Marks' VVIQ and subjects' ability to perform certain tasks that "psychologists have often supposed to require visual imagery". Such tasks include mental rotation and visual memory, both of which I can

perform easily. I find it blindingly obvious that actual visual mental imagery is not required for these tasks. Here is my own introspection about mental imagery: what I call Imagination is sufficient for all the tasks that supposedly require visual imagery.

Does anyone know of (tested?) exercises for developing visual mental imagery from scratch?

Does anyone know that developing visual mental imagery from scratch is likely impossible?

Naturalism versus unbounded (or unmaximisable) utility options

There are many paradoxes with unbounded utility functions. For instance, consider whether it's rational to spend eternity in Hell:

> Suppose that you die, and God offers you a deal. You can spend 1 day in Hell, and he will give you 2 days in Heaven, and then you will spend the rest of eternity in Purgatory (which is positioned exactly midway in utility between heaven and hell). You decide that it's a good deal, and accept. At the end of your first day in Hell, God offers you the same deal: 1 extra day in Hell, and you will get 2 more days in Heaven. Again you accept. The same deal is offered at the end of the second day.

And the result is... that you spend eternity in Hell. There is never a rational moment to leave for Heaven - that decision is always dominated by the decision to stay in Hell.

Or consider a simpler paradox:

> You're immortal. Tell Omega any natural number, and he will give you that much utility. On top of that, he will give you any utility you may have lost in the decision process (such as the time wasted choosing and specifying your number). Then he departs. What number will you choose?

Again, there's no good answer to this problem - any number you name, you could have got more by naming a higher one. And since Omega compensates you for extra effort, there's never any reason to not name a higher number.

It seems that these are problems caused by unbounded utility. But that's not the case, in fact! Consider:

> You're immortal. Tell Omega any real number r > 0, and he'll give you 1-r utility. On top of that, he will give you any utility you may have lost in the decision process (such as the time wasted choosing and specifying your number). Then he departs. What number will you choose?

Again, there is not best answer - for any r, r/2 would have been better. So these problems arise not because of unbounded utility, but because of unbounded options. You have infinitely many options to choose from (sequentially in the Heaven and Hell problem, all at once in the other two) and the set of possible utilities from your choices does not possess a maximum - so there is no best choice.

What should you do? In the Heaven and Hell problem, you end up worse off if you make the locally dominant decision at each decision node - if you always choose to add an extra day in Hell, you'll never get out of it. At some point (maybe at the very beginning), you're going to have to give up an advantageous deal. In fact, since giving up once means you'll never be offered the deal again, you're going to have to give up arbitrarily much utility. Is there a way out of this conundrum?

Assume first that you're a deterministic agent, and imagine that you're sitting down for an hour to think about this (don't worry, Satan can wait, he's just warming up the pokers). Since you're deterministic, and you know it, then your ultimate life future will be entirely determined by what you decide right now (in fact your life history is already determined, you just don't know it yet - still, by the Markov property, your current decision also determines the future). Now, you don't have to reach any grand decision now - you're just deciding what you'll do for the next hour or so. Some possible options are:

- Ignore everything, sing songs to yourself.
- Think about this some more, thinking of yourself as an algorithm.
- Think about this some more, thinking of yourself as a collection of arguing agents.
- Pick a number N, and accept all of God's deals until day N.
- Promise yourself you'll reject all of God's deals.
- Accept God's deal for today, hope something turns up.
- Defer any decision until another hour has passed.
- ...

There are many other options - in fact, there are precisely as many options as you've considered during that hour. And, crucially, you can put an estimated expected utility to each one. For instance, you might know yourself, and suspect that you'll always do the same thing (you have no self discipline where cake and Heaven are concerned), so any decision apart from immediately rejecting all of God's deals will give you $-\infty$ utility. Or maybe you know yourself, and have great self discipline and perfect precommitments- therefore if you pick a number N in the coming hour, you'll stick to it. Thinking some more may have a certain expected utility - which may differ depending on what directions you direct your thoughts. And if you know that you can't direct your thoughts - well then they'll all have the same expected utility.

But notice what's happening here: you've reduced the expected utility calculation over infinitely many options, to one over finitely many options - namely, all the interim decisions that you can consider in the course of an hour. Since you are deterministic, the infinitely many options don't have an impact: whatever interim decision you follow,

will uniquely determine how much utility you actually get out of this. And given finitely many options, each with expected utility, choosing one doesn't give any paradoxes.

And note that you don't need determinism - adding stochastic components to yourself doesn't change anything, as you're already using expected utility anyway. So all you need is an assumption of naturalism - that you're subject to the laws of nature, that your decision will be the result of deterministic or stochastic processes. In other words, you don't have 'spooky' free will that contradicts the laws of physics.

Of course, you might be wrong about your estimates - maybe you have more/less willpower than you initially thought. That doesn't invalidate the model - at every hour, at every interim decision, you need to choose the option that will, in your estimation, *ultimately* result in the most utility (not just for the next few moments or days).

If we want to be more formal, we can say that you're deciding on a decision *policy* - choosing among the different agents that you could be, the one most likely to reach high expected utility. Here are some policies you could choose from (the challenge is to find a policy that gets you the most days in Hell/Heaven, without getting stuck and going on forever):

- Decide to count the days, and reject God's deal as soon as you lose count.
- Fix a probability distribution over future days, and reject God's deal with a certain probability.
- Model yourself as a finite state machine. Figure out the Busy Beaver number of that finite state machine. Reject the deal when the number of days climbs close to that.
- Realise that you probably can't compute the Busy Beaver number for yourself, and instead use some very fast growing function like the Ackermann functions instead.
- Use the Ackermann function to count down the days *during which you formulate a policy*; after that, implement it.
- Estimate that there is a non-zero probability of falling into a loop (which would give you -∞ utility), so reject God's deal as soon as possible.
- Estimate that there is a non-zero probability of accidentally telling God the wrong thing, so commit to accepting all of God's deals (and count on accidents to rescue you from -∞ utility).

But why spend a whole hour thinking about it? Surely the same applies for half an hour, a minute, a second, a microsecond? That's entirely a convenience choice - if you think about things in one second increments, then the interim decision "think some more" is nearly always going to be the dominant one.

The mention of the Busy Beaver number hints at a truth - given the limitations of your mind and decision abilities, there is one policy, among all possible policies that you could implement, that gives you the most utility. More complicated policies you can't implement (which generally means you'd hit a loop and get -∞ utility), and simpler policies would give you less utility. Of course, you likely won't find that policy, or anything close to it. It all really depends on how good your policy finding policy is (and your policy finding policy finding policy...).

That's maybe the most important aspect of these problems: some agents are just better than others. Unlike finite cases where any agent can simply list all the options, take their time, and choose the best one, here an agent with a better decision algorithm will outperform another. Even if they start with the same resources (memory capacity, cognitive shortcuts, etc...) one may be a lot better than another. If the agents

don't acquire more resources during their time in Hell, then their maximal possible utility is related to their Busy Beaver number - basically the maximal length that a finite-state agent can survive without falling into an infinite loop. Busy Beaver numbers are extremely uncomputable, so some agents, by pure chance, may be capable of acquiring much greater utility than others. And agents that start with more resources have a much larger theoretical maximum - not fair, but deal with it. Hence it's not really an infinite option scenario, but an infinite agent scenario, with each agent having a different maximal expected utility that they can extract from the setup.

It should be noted that God, or any being capable of hypercomputation, has real problems in these situations: they actually have infinite options (not a finite options of choosing their future policy), and so don't have any solution available.

This is also related to theoretical maximally optimum agent that is AIXI: for any computable agent that approximates AIXI, there will be other agents that approximate it better (and hence get higher expected utility). Again, it's not fair, but not unexpected either: smarter agents are smarter.

What to do?

This analysis doesn't solve the vexing question of what to do - what is the right answer to these kind of problems? These depend on what type of agent you are, but what you need to do is estimate the maximal integer you are capable of computing (and storing), and endure for that many days. Certain probabilistic strategies may improve your performance further, but you have to put the effort into finding them.

Learning critical thinking: a personal example

Related to: Is Rationality Teachable

“Critical care nursing isn’t about having critically ill patients,” my preceptor likes to say, “it’s about critical thinking.”

I doubt she's talking about the same kind of critical thinking that philosophers are, and I find that definition abstract anyway. There’s been a lot of talk about critical thinking during our four years of nursing school, but our profs seem to have a hard time defining it. So I’ll go with a definition from Google.

> *Critical thinking can be seen as having two components: 1) a set of information and belief generating and processing skills, and 2) the habit, based on intellectual commitment, of using those skills to guide behaviour. It is thus to be contrasted with: 1) the mere acquisition and retention of information alone, because it involves a particular way in which information is sought and treated; 2) the mere possession of a set of skills, because it involves the continual use of them; and 3) the mere use of those skills ("as an exercise") without acceptance of their results.*[1]

That’s basically rationality–epistemic, i.e. generating true beliefs, and instrumental, i.e. knowing how to use them to achieve what you want. Maybe part of me expected, implicitly, to have an easier time learning this skill because of my Less Wrong knowledge. And maybe I am more consciously aware of my mistakes, and the cognitive factors that caused them, than most of my classmates. When it’s forty-five minutes past the end of my shift and I’m still charting, I’m also calling myself out on succumbing to the planning fallacy. I once went through the first half hour of a shift during my pediatrics rotation thinking that one of my patients had cerebral palsy, when he actually had cystic fibrosis–all because I misread my prof’s handwriting as ‘CP’ when she’d written ‘CF’. I was totally confused by all the enzyme supplements on his list of meds, but it still took me a while to figure it out–a combination of priming and confirmation bias, taken to the next level.

But, overall, even if I know what I'm doing wrong, it *hasn’t* been easier to do things right. I have a hard time with the hospital environment, possibly because I’m the kind of person who ended up reading and posting on Less Wrong. My cognitive style leans towards Type 2 reasoning, in Keith Stanovich’s taxonomy–thorough, but slow. I like to *understand* things, on a deep level. I like knowing why I’m doing something, and I don’t trust my intuitions, the fast-and-dirty product of Type 1 reasoning. But Type 2 reasoning requires a lot of working memory, and humans aren’t known for that, which is the source of most of my frustration and nearly all of my errors–when working memory overload forces me to be a cognitive miser.

Still, for all the frustration, I’m pretty sure I’ve ended up in the *perfect* environment to learn this skill called ‘critical thinking.’ I’m way out of my depth–which I expected. *No* fourth year student is ready to work independently in a trauma ICU, but I decided to finish my schooling here in the name of tsuyoku naritai, and for all the days when I’ve gone home crying, it’s still worth it. I’m *learning*.

The skills

1. *A set of information and belief generating and processing skills.*

Medicine, and nursing, are a bit like physics, in that you need to generate true beliefs about systems that exist outside of you, and predict how they're going to behave. This involves knowing a lot of abstract theory, which I'm good at, and a lot of heuristics and pattern-matching for applying the right bits of theory to particular patients, which I'm less good at. That's partly an experience thing; my brain needs patterns to match to. But in general, I have decent mental models of my patients. I'm curious and I like to understand things. If I don't know what part of the theories applies, I *ask*.

2. *The habit, based on intellectual commitment, of using those skills to guide behaviour.*

So you've got your mental model of your patient, your best understand of what's actually going on, on a physiological and biochemical level, down under the skin where you can't see it. You know what "normal" is for a variety of measures: vital signs, lung sounds, lab values, etc. Given that your patient is in the ICU, you know *something's* abnormal, or they wouldn't be there. Their diagnosis tells you what to expect, and you look at the results of your assessments and ask a couple of questions. One: is this what I expect, for this patient? Two: what do I need to do about it?

I'm not going to be *surprised* if a post-op patient has low hemoglobin. It's information of a kind, telling the doctor whether or not the patient needs a transfusion, and how many units, but it's not really new information, and a moderately abnormal value wouldn't worry me or anyone else. If their hemoglobin *keeps* dropping; okay, they're actively bleeding somewhere, that's irritating, and possibly dangerous, and needs dealing with, but it's not surprising.

But if a patient here for an abdominal surgery suddenly has decreased level of consciousness and their pupils aren't reacting normally to light, I'm *worried*. There's nothing in my mental model that says I should expect it. I notice I'm confused, and that confusion guides my behaviour; I call the doctor right away, because we need more information to update our collective mental model, information you can't get just from observation, like a CT scan of the head. (Even this is optimistic–plenty of patients are admitted to the ICU because we have no idea what's wrong with them, and are hoping to keep them alive long enough to find out.)

The basics of ICU nursing come down to treating numbers. Heart rate, blood pressure, oxygen saturations, urine output, etc; know the acceptable range, notice if they change, and use Treatment X to get them back where they're supposed to be. Which doesn't sound that hard. But implicit in 'notice if they change' is 'figure out *why* they changed', because that affects how you treat them, and implicit in that is a *lot* of background knowledge, which has to be put in context.

I'm, honestly, fairly terrible at this. It's a compartmentalization thing. I don't like using my knowledge as input arguments to generate new conclusions and then relying on those conclusions to treat human beings. It feels like *guessing*. Even though, back in high school, I never really needed to study for physics tests–if I *understood* what we'd learned, I could re-derive forgotten details from first principles. But hospital patients ended up in a non-overlapping magisterium in my head. In order for me to trust my knowledge, it has to have come *directly* from the lips of a teacher or experienced nurse.

My preceptor, who hates this. “She needs to continue to work on her critical thinking when it comes to caring for critically ill patients,” she wrote on my evaluation. “She knows the theory, and is now working to apply it to ICU nursing.” Shorthand for, *she knows the theory, but getting her to apply it to ICU nursing is like pulling teeth*. A number of our conversations have gone like this:

> Me: “Our patient’s blood pressure dropped a bit.”
>
> Her: “Yeah, it did. What do you want to do about it?”
>
> Me: “I, uh, I don’t know... Should I increase the vasopressors?”
>
> Her: “I don’t know, should you?”
>
> Me: “Uh, maybe I should increase the phenylephrine to 40 mcg/min and see what happens. How long should I wait to see?”
>
> Her: “You tell me.”
>
> Me: “Well, let’s say it’ll take a few minutes for what’s in the tubing now to get pushed through, and it should take effect pretty quickly because it’s IV, like a minute... So if his blood pressure’s not up enough in five minutes, I’ll increase the phenyl to 60. Does that sound okay?”
>
> Her: “It’s your decision to make."

Needless to say, I find this teaching method extremely stressful and scary, and I’m learning about ten times more than I would if she *answered* the questions I asked. Because “the mere acquisition and retention of information alone” isn’t my problem. I have a brain like an encyclopaedia. My problem, in the critical care nursing context, is the “particular way in which information is sought and treated.” I need to know the right time to notice something is wrong, the right place to look in my encyclopaedia, and the right way to take the information I just looked up and figure out what to *do* with it.

The mistakes

Some of my errors, unsurprisingly, boil down to a failure to override inappropriate Type 1 responses with Type 2 responses–in other words, not thinking about what I’m doing. But *most* of them are more of a mindware gap–I don’t yet have the “domain-specific knowledge sets” that the nurses around me have. Not just theory knowledge; I do have *most* of that; but the procedural habits of how to stay organized and prioritize and dump the contents of my working memory onto paper in a way that I can read them back later. Usually, when I make a mistake, I knew better, but the part of my brain that knew better was doing something else at the time, that small note of confusion getting lost in the general chaos.

Pretty much all nurses keep a “feuille de route”–I have yet to find a satisfactory English word for this, but it’s a personal sheet of paper, *not* legal charting, usually kept in a pocket, and used as an extended working memory. In med/surg, when I had four patients, I made a chart with four columns; name and personal information, medications, treatments/general plan for the day, and medical history; and as many rows as I had patients. If something was important, I circled it in red ink. This system

doesn’t work in the ICU, so my current feuille de route has several aspects. I fold a piece of blank paper into four, and take notes from the previous shift report on one quarter of one side, or two quarters if it’s a long report. Across from that, I draw a vertical column of times, from 8:00 am to 6:00 pm (or 8:00 pm to 6:00 am). 7:00 pm and 7:00 am are shift change, so nothing else really gets done for that hour. I use this to scribble down what I need to get down during my twelve hours, and approximately when I want to do it, and I prioritize, i.e. from 1 to 5 most to least important. Once it’s done, I cross it off–then I can forget about it. On the other side of the paper, I make a cheat sheet for giving report to the next nurse, or presenting my patient to the doctors at rounds.

This might be low-tech and simple, but it takes a *huge* load off my working memory, and reduces my most frequent error, which is to get so overwhelmed and frazzled that my brain goes on strike. In other words, the failure to override Type 1 responses due to the lack of cognitive capacity to run a Type 2 process. It’s drastically cut down on the frequency of this mental conversation:

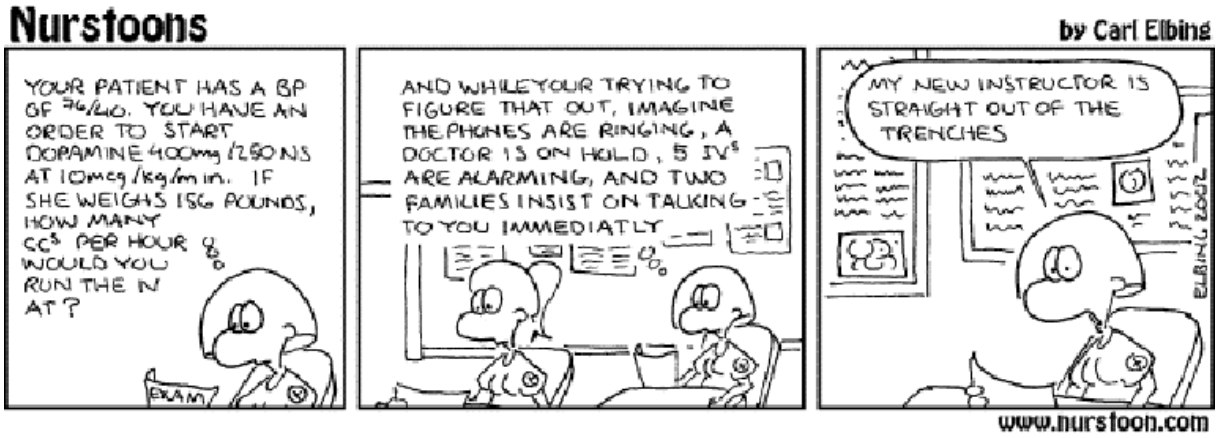
> Me: “I turned off the sedation, and my patient isn’t waking up as fast as I expected. I notice I’m confused–”
>
> My brain: “You’re always confused! Everything around here is intensely confusing! How am I supposed to use that as information?”

Odd as it might sound, I often don’t *notice* when my brain starts edging towards a meltdown. The feeling itself is quite recognizable, but the circumstances that lead to it, i.e. overloaded working memory, mean that I’m not usually paying attention to my own feelings.

“You need to stop and take a breath,” my preceptor says about fifty times a day. Easier said than done–but it’s more efficient, overall, to have a tiny part of my mind permanently on standby, keeping an eye on my emotions, noticing when the gears start to overheat. Then stop, take a breath, and let go of *everything* except the task at hand, trusting myself to have created enough cues in my environment to retrieve the other tasks, once I’m done. Humans don’t multitask well. Doing one thing while trying to remember a list of five others is intense multitasking, and it’s no wonder it’s exhausting.

The implications

“You can’t teach critical thinking,” my preceptor says, but I’m pretty sure that’s exactly what she’s doing right now. A great deal of what I already know is domain-specific to nursing, but most of what I’m learning right now is generally applicable. I’m learning the procedural skills to work through difficult problems, under what Keith Stanovich would call average rather than optimal conditions. Sitting in my own little bubble in front of a multiple choice exam–that’s optimal conditions. Trying to figure out if I should be surprised or worried about my patient’s increased heart rate, while simultaneously deciding whether or not I can ignore the ventilator alarm and whether I can finish giving my twelve o’clock antibiotic before I need to do twelve o’clock vitals–that’s not just average conditions, it’s under-duress conditions.



I’m hoping that after a few more weeks, or maybe a few more years, I’ll be able to perform comfortably in this intensely terrifying environment. And I’m hoping that some of the skills I learn will be general-purpose, for me at least. It’d be nice if they were teachable to others, too, but I think my preceptor might be right about one thing–you can’t teach this kind of critical thinking in the classroom. It's about moulding my brain into the right shape, and everyone's brain starts out in a different shape, so the mould has to be personalized.

But the habits are general ones. Notice when you're faced with a difficult problem, or making an important decision. Notice that you're doing this while distracted. Stop and take a breath. Get out a piece of paper. Figure out how the problem is formatted in your mind, and format it that way on the paper. (This is probably the hardest part). Dump your working memory and give yourself space to think. Prioritize from 1 to *n*. Keep an eye on the evolving situation, sure, but find that moment of concentration in the midst of chaos, and solve the problem.

Of course, it's far from guaranteed that this will work. I'm making an empirical prediction; that the skills I'm currently learning *will* be transferable to non-nursing areas, and that they'll make a difference in my life outside of work. I'll be on the lookout for examples, either of success or failure.

References

Scriven, Michael; Paul, Richard. *Defining critical thinking.* (2011). The critical thinking community. *http://www.criticalthinking.org/pages/defining-critical-thinking/410*

Memetic Tribalism

Related: politics is the mind killer, other optimizing

When someone says something stupid, I get an urge to correct them. Based on the stories I hear from others, I'm not the only one.

For example, some of my friends are into this rationality thing, and they've learned about all these biases and correct ways to get things done. Naturally, they get irritated with people who haven't learned this stuff. They complain about how their family members or coworkers aren't rational, and they ask what is the best way to correct them.

I could get into the details of the optimal set of arguments to turn someone into a rationalist, or I could go a bit meta and ask: "Why would you want to do that?"

Why should you spend your time correcting someone else's reasoning?

One reason that comes up is that it's valuable for some reason to change their reasoning. OK, when is it possible?

1. You actually know better than them.
2. You know how to patch their reasoning.
3. They will be receptive to said patching.
4. They will actually change their behavior if the accept the patch.

It seems like it should be rather rare for those conditions to all be true, or even to be likely enough for the expected gain to be worth the cost, and yet I feel the urge quite often. And I'm not thinking it through and deciding, I'm just feeling an urge; humans are adaptation executors, and this one seems like an adaptation. For some reason "correcting" people's reasoning was important enough in the ancestral environment to be special-cased in motivation hardware.

I could try to spin an ev-psych just-so story about tribal status, intellectual dominance hierarchies, ingroup-outgroup signaling, and whatnot, but I'm not an evolutionary psychologist, so I wouldn't actually know what I was doing, and the details don't matter anyway. What matters is that this urge seems to be hardware, and it probably has nothing to do with actual truth or *your* strategic concerns.

It seems to happen to everyone who has ideas. Social justice types get frustrated with people who seem unable to acknowledge their own privilege. The epistemological flamewar between atheists and theists rages continually across the internet. Tech-savvy folk get frustrated with others' total inability to explore and use Google. Some aspiring rationalists get annoyed with people who refuse to decompartmentalize or claim that something is in a separate magisteria.

Some of those border on being just classic blue vs green thinking, but from the outside, the rationality example isn't all that different. They all seem to be motivated mostly by "This person fails to display the complex habits of thought that I think are fashionable; I should {make fun | correct them | call them out}."

I'm now quite skeptical that my urge to correct reflects an actual opportunity to win by improving someone's thinking, given that I'd feel it whether or not I could actually help, and that it seems to be caused by something else.

The value of attempting a rationality-intervention has gone back down towards baseline, but it's not obvious that the baseline value of rationality interventions is all that low. Maybe it's a good idea, even if there is a possible bias supporting it. We can't win just by reversing our biases; reversed stupidity is not intelligence.

The best reason I can think of to correct flawed thinking is if your ability to accomplish your goals directly depends on their rationality. Maybe they are your business partner, or your spouse. Someone specific and close who you can cooperate with a lot. If this is the case, it's near the same level of urgency as correcting your own.

Another good reason (to discuss the subject at least) is that discussing your ideas with smart people is a good way to make your ideas better. I often get my dad to poke holes in my current craziness, because he is smarter and wiser than me. If this is your angle, keep in mind that if you expect someone else to correct you, it's probably not best to go in making bold claims and implicitly claiming intellectual dominance.

An OK reason is that creating more rationalists is valuable in general. This one is less good than it first appears. Do you really think your comparative advantage right now is in converting this person to your way of thinking? Is that really worth the risk of social friction and expenditure of time and mental energy? *Is this the best method you can think of for creating more rationalists?*

I think it is valuable to raise the sanity waterline when you can, but using methods of mass instruction like writing blog posts, administering a meetup, or launching a whole rationality movement is a lot more effective than arguing with your mom. Those options aren't for everybody of course, but if you're into waterline-manipulation, you should at least be considering strategies like them. At least consider picking a better time.

Another reason that gets brought up is that turning people around you into rationalists is instrumental in a selfish way, because it makes life easier for you. This one is suspect to me, even without the incentive to rationalize. Did you also seriously consider *sabotaging* people's rationality to take advantage of them? Surely that's nearly as plausible a-priori. For what specific reason did your search process rank cooperation over predation?

I'm sure there are plenty of good reasons to prefer cooperation, but of course no search process was ever run. All of these reasons that come to mind when I think of why I might want to fix someone's reasoning are just post-hoc rationalizations of an automatic behavior. The true chain of cause-and-effect is observe->feel->act; no planning or thinking involved, except where it is necessary for the act. And that feeling isn't specific to rationality, it affects all mental habits, even stupid ones.

Rationality isn't just a new memetic orthodoxy for the cool kids, it's about actually winning. Every improvement requires a change. Rationalizing strategic reasons for instinctual behavior isn't change, it's spending your resources answering questions with *zero value of information*. Rationality isn't about what other people are doing wrong; it's about what *you* are doing wrong.

I used to call this practice of modeling other people's thoughts to enforce orthodoxy on them "incorrect use of empathy", but in terms of ev-psych, it may be exactly the

correct use of empathy. We can call it Memetic Tribalism instead.

(I've ignored the *other* reason to correct people's reasoning, which is that it's fun and status-increasing. When I reflect on my reasons for writing posts like this, it turns out I do it largely for the fun and internet status points, but I try to at least be aware of that.)