Fun Theory

Eliezer Yudkowsky

2008 - 2010

Contents

31 Laws of Fun

So this is Utopia, is it? Well
I beg your pardon, I thought it was Hell.
— Sir Max Beerholm, verse entitled
In a Copy of More's (or Shaw's or Wells's or Plato's or Anybody's) Utopia

This is a shorter summary of the Fun Theory Sequence with all the background theory left out - just the compressed advice to the would-be author or futurist who wishes to imagine a world where people *might actually want to live:*

1. Think of a *typical day* in the life of someone who's been adapting to Utopia *for a while*. Don't anchor on the first moment of "hearing the good news". Heaven's "You'll never have to work again, and the streets are paved with gold!" sounds like good news to a tired and poverty-stricken peasant, but two months later it might not be so much fun. (Prolegomena to a Theory of Fun.)

2. Beware of packing your Utopia with things you think people *should* do that aren't actually *fun*. Again, consider Christian Heaven: singing hymns doesn't sound like loads of endless fun, but you're *supposed*to enjoy praying, so no one can point this out. (Prolegomena to a Theory of Fun.)

3. Making a video game easier doesn't always improve it. The same holds true of a life. Think in terms of clearing out low-quality drudgery to make way for high-quality challenge, rather than eliminating work. (High Challenge.)

4. Life should contain novelty - experiences you haven't encountered before, preferably teaching you something you didn't already know. If there isn't a sufficient supply of novelty (relative to the speed at which you generalize), you'll get bored. (Complex Novelty.)

5. People should get smarter at a rate sufficient to integrate their old experiences, but not so much smarter so fast that they can't integrate their new intelligence. Being smarter means you get bored faster, but you can also tackle new challenges you couldn't understand before. (Complex Novelty.)

6. People should live in a world that fully engages their senses, their bodies, and their brains. This means either that the world resembles the ancestral savanna more than say a windowless office; or alternatively, that brains and bodies have changed to be fully engaged by different kinds of complicated challenges and environments. (Fictions intended to entertain a human audience should concentrate primarily on the former option.) (Sensual Experience.)

7. Timothy Ferris: "What is the opposite of happiness? Sadness? No. Just as love and hate are two sides of the same coin, so are happiness and sadness... The opposite of love is indifference, and the opposite of happiness is - here's the clincher - boredom... The question you should be asking isn't 'What do I want?' or 'What are my goals?' but 'What would excite me?'... *Living* like a millionaire requires *doing* interesting things and not just owning enviable things." (Existential Angst Factory.)

8. Any particular individual's life should get better and better over time. (Continuous Improvement.)

9. You should not know exactly what improvements the future holds, although you should look forward to finding out. The actual event should come as a pleasant surprise. (Justified Expectation of Pleasant Surprises.)

10. Our hunter-gatherer ancestors strung their own bows, wove their own baskets and whittled their own flutes; then they did their own hunting, their own gathering and played their own music. Futuristic Utopias are often depicted as offering more and more neat buttons that do less and less comprehensible things *for* you. Ask not what interesting things Utopia can do *for* people; ask rather what interesting things the inhabitants could do for *themselves*- with their own brains, their own bodies, or tools they understand how to build. (Living By Your Own Strength.)

11. Living in Eutopia should make people stronger, not weaker, over time. The inhabitants should appear *more formidable* than the people of our own world, not less. (Living By Your Own Strength; see also, Tsuyoku Naritai.)

12. Life should not be broken up into a series of disconnected episodes with no long-term consequences. No matter how sensual or complex, playing one *really great video game* after another, does not make a life story. (Emotional Involvement.)

13. People should make their own destinies; their lives should not be choreographed to the point that they no longer need to imagine, plan and navigate their own futures. Citizens should not be the pawns of more powerful gods, still less their sculpted material. One simple solution would be to have the world work by stable rules that are the same for everyone, where the burden of Eutopia is carried by a good initial choice of rules, rather than by any optimization pressure applied to individual lives. (Free to Optimize.)

14. Human minds should not have to *play on a level field* with vastly superior entities. Most people don't like being overshadowed. Gods destroy a human protagonist's "main character" status; this is undesirable in fiction and probably in real life. (E.g.: C. S. Lewis's Narnia, Iain Banks's Culture.) Either change people's emotional makeup so that they don't *mind* being unnecessary, or keep the gods *way* off their playing field. Fictional stories intended for human audiences cannot do the former. (And in real life, you probably *can* have powerful AIs that are neither sentient nor meddlesome. See the main post and its prerequisites.) (Amputation of Destiny.)

15. Trying to compete on a single flat playing field with *six billion other humans* also creates problems. Our ancestors lived in bands of around 50 people. Today the media is constantly bombarding us with news of *exceptionally* rich and pretty people as if they lived next door to us; and very few people get a chance to be the *best* at any specialty. (Dunbar's Function.)

16. Our ancestors also had some degree of genuine control over their band's politics. Contrast to modern nation-states where almost no one knows

the President on a personal level or could argue Congress out of a bad decision. (Though that doesn't stop people from arguing as loudly as if they still lived in a 50-person band.) (Dunbar's Function.)

17. Offering people more options is *not*always helping them (especially if the option is something they couldn't do for themselves). Losses are more painful than the corresponding gains, so if choices are different along many dimensions and only one choice can be taken, people tend to focus on the loss of the road *not*taken. Offering a road that bypasses a challenge makes the challenge feel less real, even if the cheat is diligently refused. It is also a sad fact that humans predictably make certain kinds of mistakes. Don't assume that building *more choice* into your Utopia is necessarily an improvement because "people can always just say no". This *sounds reassuring* to an outside reader - "Don't worry, *you'll* decide! You trust *yourself*, right?" - but might not be much fun to actually *live*with. (Harmful Options.)

18. Extreme example of the above: being constantly offered huge temptations that are incredibly dangerous - a completely realistic virtual world, or very addictive and pleasurable drugs. You can never allow yourself a single moment of willpower failure over your whole life. (E.g.: John C. Wright's Golden Oecumene.) (Devil's Offers.)

19. Conversely, when people are grown strong enough to shoot off their feet *without external help,* stopping them may be too much interference. Hopefully they'll then be smart enough *not* to: By the time they can build the gun, they'll know what happens if they pull the gun, and won't need a smothering safety blanket. If that's the theory, then dangerous options need correspondingly difficult locks. (Devil's Offers.)

20. *Telling* people truths they haven't yet figured out for themselves, is not always helping them. (Joy in Discovery.)

21. Brains are some of the most complicated things in the world. Thus, other humans (other minds) are some of the most complicated things we deal with. For us, this interaction has a unique character because of the *sympathy*we feel for others - the way that our brain tends to align with their brain - rather than our brain just treating other brains as big complicated machines with levers to pull. Reducing the need for people to interact with other people reduces the complexity of human existence; this is a step in the wrong direction. For example, resist the temptation to *simplify people's lives* by offering them artificially perfect sexual/romantic partners. (Interpersonal Entanglement.)

22. But admittedly, humanity does have a *statistical*sex problem: the male distribution of attributes doesn't harmonize with the female distribution of desires, or vice versa. Not everything in Eutopia should be easy - but it shouldn't be pointlessly, unresolvably frustrating either. (This is

a general principle.) So imagine *nudging* the distributions to make the problem *solvable* - rather than waving a magic wand and solving everything instantly. (Interpersonal Entanglement.)

23. In general, tampering with brains, minds, emotions, and personalities is way more fraught on every possible level of ethics and difficulty, than tampering with bodies and environments. Always ask what you can do by messing with the environment before you imagine messing with minds. Then prefer small cognitive changes to big ones. You're not just outrunning your human audience, you're outrunning your own imagination. (Changing Emotions.)

24. In this present world, there is an imbalance between pleasure and pain. An unskilled torturer with simple tools can create worse pain in thirty seconds, than an extremely skilled sexual artist can create pleasure in thirty minutes. One response would be to *remedy the imbalance* - to have the world contain *more*joy than sorrow. Pain might exist, but not pointless endless unendurable pain. Mistakes would have more *proportionate*penalties: You might touch a hot stove and end up with a painful blister; but not glance away for two seconds and spend the rest of your life in a wheelchair. The people would be stronger, less exhausted. This path would eliminate *mind-destroying* pain, and make pleasure more abundant. Another path would eliminate pain *entirely*. Whatever the relative merits of the real-world proposals, *fictional stories cannot take the second path.* (Serious Stories.)

25. George Orwell once observed that Utopias are chiefly concerned with avoiding fuss. Don't be afraid to write a loud Eutopia that might wake up the neighbors. (Eutopia is Scary; George Orwell's Why Socialists Don't Believe in Fun.)

26. George Orwell observed that "The inhabitants of perfect universes seem to have no spontaneous gaiety and are usually somewhat repulsive into the bargain." If you write a story and your characters turn out like this, it probably reflects some much deeper flaw that can't be fixed by having the State hire a few clowns. (George Orwell's Why Socialists Don't Believe in Fun.)

27. Ben Franklin, yanked into our own era, would be surprised and delighted by some aspects of his Future. Other aspects would horrify, disgust, and *frighten* him; and this is not because our world has gone *wrong*, but because it has *improved*relative to his time. Relatively few things would have gone *just*as Ben Franklin expected. If you imagine a world which your imagination finds familiar and comforting, it will inspire few others, and the whole exercise will lack integrity. Try to conceive of a genuinely better world in which you, yourself, would be *shocked*(at least at first) and *out of place* (at least at first). (Eutopia is Scary.)

28. Utopia and Dystopia are two sides of the same coin; both just confirm the moral sensibilities you started with. Whether the world is a libertarian utopia of government non-interference, or a hellish dystopia of government intrusion and regulation, you get to say "I was right all along." Don't just imagine something that conforms to your *existing*ideals of government, relationships, politics, work, or daily life. Find the better world that zogs instead of zigging or zagging. (To safeguard your sensibilities, you can tell yourself it's just an *arguably* better world but isn't *really* better than your favorite standard Utopia... but you'll know you're *really*doing it right if you find your ideals *changing*.) (Building Weirdtopia.)

29. If your Utopia still seems like an endless gloomy drudgery of existential angst no matter how much you try to brighten it, there's at least one major problem that you're *entirely failing to focus on*. (Existential Angst Factory.)

30. 'Tis a sad mind that cares about nothing except itself. In the modern-day world, if an altruist looks around, their eye is caught by large groups of people in desperate jeopardy. People in a better world will *not*see this: A true Eutopia will run low on victims to be rescued. This doesn't imply that the inhabitants look around outside themselves and see *nothing*. They may care about friends and family, truth and freedom, common projects; outside minds, shared goals, and high ideals. (Higher Purpose.)

31. Still, a story that confronts the challenge of Eutopia should *not*just have the convenient plot of "The Dark Lord Sauron is about to invade and kill everybody". The would-be author will have to find something *slightly less awful*for his characters to *legitimately care about*. This is part of the challenge of showing that human progress is not the end of human stories, and that people *not* in imminent danger of death can still lead interesting lives. Those of you interested in confronting lethal planetary-sized dangers should focus on *present-day real life*. (Higher Purpose.)

The simultaneous solution of all these design requirements is left as an exercise to the reader. At least for now.

The enumeration in this post of certain Laws shall not be construed to deny or disparage others not mentioned. I didn't happen to write about humor, but it would be a sad world that held no laughter, etcetera.

To anyone seriously interested in trying to write a Eutopian story using these Laws: You must first know *how to write*. There are many, many books on how to write; you should read at least three; and they will all tell you that a great deal of practice is required. Your practice stories should *not*be composed anywhere so difficult as Eutopia. That said, my *second*most important advice for authors is this: Life will never become boringly easy for your characters so long as they can make things difficult for each other.

Finally, this dire warning: Concretely imagining worlds much better than your present-day real life, may suck out your soul like an emotional vacuum cleaner. (See Seduced by Imagination.) Fun Theory is dangerous, use it with caution, you have been warned.

Prolegomena to a Theory of Fun

Raise the topic of cryonics, uploading, or just medically extended lifespan/healthspan, and some bioconservative neo-Luddite is bound to ask, in portentous tones:

"But what will people *do* all day?"

They don't try to actually answer the question. That is not a bioethicist's role, in the scheme of things. They're just there to collect credit for the Deep Wisdom of asking the question. It's enough to *imply* that the question is unanswerable, and therefore, we should all drop dead.

That doesn't mean it's a *bad* question.

It's not an *easy* question to answer, either. The primary experimental result in hedonic psychology - the study of happiness - is that people don't *know* what makes them happy.

And there are many exciting results in this new field, which go a long way toward explaining the emptiness of classical Utopias. But it's worth remembering that *human* hedonic psychology is not enough for us to consider, if we're asking whether a million-year lifespan could be worth living.

Fun Theory, then, is the field of knowledge that would deal in questions like:

- "How much fun is there in the universe?"
- "Will we ever run out of fun?"
- "Are we having fun yet?"
- "Could we be having more fun?"

One major set of experimental results in hedonic psychology has to do with *overestimating the impact* of life events on happiness. Six months after the event, lottery winners aren't as happy as they expected to be, and quadriplegics aren't as sad. A parent who loses a child isn't as sad as they think they'll be, a few years later. If you look at one moment snapshotted out of their lives a few years later, that moment isn't likely to be about the lost child. Maybe they're playing with one of their surviving children on a swing. Maybe they're just listening to a nice song on the radio.

When people are asked to imagine how happy or sad an event will make them, they anchor on *the moment of first receiving the news,* rather than realistically imagining the process of daily life years later.

Consider what the Christians made of their Heaven, meant to be literally *eternal.* Endless rest, the glorious presence of God, and occasionally - in the more clueless sort of sermon - golden streets and diamond buildings. Is this eudaimonia? It doesn't even seem very *hedonic.*

As someone who said his share of prayers back in his Orthodox Jewish childhood upbringing, I can personally testify that praising God is an enormously boring activity, even if you're still young enough to truly believe in God. The part about praising God is there as an applause light that no one is allowed to contradict: it's something theists believe they *should* enjoy, even though, if you ran them through an fMRI machine, you probably wouldn't find their pleasure centers lighting up much.

Ideology is one major wellspring of flawed Utopias, containing things that the imaginer believes *should*be enjoyed, rather than things that would actually be enjoyable.

And eternal *rest?* What could possibly be more boring than eternal *rest?*

But to an exhausted, poverty-stricken medieval peasant, the Christian Heaven sounds like *good news in the moment of being first informed:* You can lay down the plow and rest! Forever! Never to work again!

It'd get boring after... what, a week? A day? An hour?

Heaven is not configured as a nice place to *live.* It is rather memetically optimized to be a nice place for an exhausted peasant to *imagine.* It's not like some Christians *actually*got a chance to live in various Heavens, and voted on how well they liked it after a year, and then they kept the best one. The Paradise that survived was the one that was *retold,* not lived.

Timothy Feriss observed, "*Living* like a millionaire requires *doing* interesting things and not just owning enviable things." Golden streets and diamond walls would fade swiftly into the background, once *obtained*- but so long as you don't actually *have* gold, it stays desirable.

And there's two lessons required to get past such failures; and these lessons are in some sense opposite to one another.

The first lesson is that humans are terrible judges of what will *actually*make them happy, in the real world and the living moments. Daniel Gilbert's *Stumbling on Happiness* is the most famous popular introduction to the research.

We need to be ready to correct for such biases - the world that is fun to *live in*, may not be the world that sounds good when spoken into our ears.

And the second lesson is that there's *nothing* in the universe out of which to construct Fun Theory, except that which we want for ourselves or prefer to become.

If, *in fact*, you *don't* like praying, then there's no higher God than yourself to tell you that you *should* enjoy it. We sometimes do things we don't like, but that's still our own choice. There's no *outside* force to scold us for making the wrong decision.

This is something for transhumanists to keep in mind - not because we're tempted to pray, of course, but because there are so many other logical-sounding solutions we wouldn't really *want.*

The transhumanist philosopher David Pearce is an advocate of what he calls the Hedonistic Imperative: The eudaimonic life is the one that is as pleasurable as possible. So even happiness attained through drugs is good? Yes, in fact: Pearce's motto is "Better Living Through Chemistry".

Or similarly: When giving a small informal talk once on the Stanford campus, I raised the topic of Fun Theory in the post-talk mingling. And someone there said that his ultimate objective was to experience delta pleasure. That's "delta" as in the Dirac delta - roughly, an infinitely high spike (that happens to be integrable). "Why?" I asked. He said, "Because that means I win."

(I replied, "How about if you get two times delta pleasure? Do you win twice as hard?")

In the transhumanist lexicon, "orgasmium" refers to simplified brains that are just pleasure centers experiencing huge amounts of stimulation - a happiness counter containing a large number, plus whatever the minimum surrounding framework to *experience* it. You can imagine a whole galaxy tiled with orgasmium. Would this be a good thing?

And the vertigo-inducing thought is this - if you would *prefer* not to become orgasmium, then why *should* you?

Mind you, there are many reasons why something that sounds unpreferred at first glance, might be worth a closer look. That was the *first* lesson. Many Christians *think*they want to go to Heaven.

But when it comes to the question, "Don't I *have* to want to be as happy as possible?" then the answer is simply "No. If you don't prefer it, why go there?"

There's nothing *except*such preferences out of which to construct Fun Theory - a second look is still a look, and must still be constructed out of preferences at some level.

In the era of my foolish youth, when I went into an affective death spiral around intelligence, I thought that the mysterious "right" thing that any superintelligence would inevitably do, would be to upgrade every nearby mind to superintelligence as fast as possible. Intelligence was good; therefore, more intelligence was better.

Somewhat later I imagined the scenario of *unlimited* computing power, so that no matter how smart you got, you were still just as far from infinity as

ever. That got me thinking about a journey rather than a destination, and *allowed* me to think "What *rate* of intelligence increase would be fun?"

But the real break came when I naturalized my understanding of morality, and value stopped being a mysterious attribute of unknown origins.

Then if there was no outside light in the sky to order me to do things -

The thought occurred to me that I didn't actually *want* to bloat up immediately into a superintelligence, *or* have my world transformed instantaneously and completely into something incomprehensible. I'd prefer to have it happen gradually, with time to stop and smell the flowers along the way.

It felt like a very guilty thought, but -

But there was nothing *higher* to *override* this preference.

In which case, if the Friendly AI project succeeded, there would be a day after the Singularity to wake up to, and myself to wake up to it.

You may not see why this would be a vertigo-inducing concept. Pretend you're Eliezer_{2003} who has spent the last seven years talking about how it's forbidden to try to look beyond the Singularity - because the AI is smarter than you, and if you knew what it would do, you would have to be that smart yourself -

- but what if you don't *want* the world to be made suddenly incomprehensible? Then there might be something to understand, that next morning, *because* you don't *actually want* to wake up in an incomprehensible world, any more than you *actually want* to suddenly be a superintelligence, or turn into orgasmium.

I can only analogize the experience to a theist who's suddenly told that they *can* know the mind of God, and it turns out to be only twenty lines of Python.

You may find it hard to sympathize. Well, Eliezer_{1996}, who originally made the mistake, was smart but methodologically inept, as I've mentioned a few times.

Still, expect to see some outraged comments on this very blog post, from commenters who think that it's *selfish and immoral*, and above all a *failure of imagination,* to talk about human-level minds still running around the day after the Singularity.

That's the frame of mind I used to occupy - that the things I wanted were selfish, and that I shouldn't think about them too much, or at all, because I would need to sacrifice them for something higher.

People who talk about an existential pit of meaninglessness in a universe devoid of meaning - I'm pretty sure they don't understand morality in naturalistic terms. There *is* vertigo involved, but it's *not* the vertigo of meaninglessness.

More like a theist who is frightened that someday God will order him to murder children, and then he realizes that there *is* no God and his fear of being ordered

to murder children *was morality.* It's a strange relief, mixed with the realization that you've been very silly, as the last remnant of outrage at your own selfishness fades away.

So the first step toward Fun Theory is that, so far as I can tell, it looks basically *okay*to make our future light cone - all the galaxies that we can get our hands on - into a place that is *fun* rather than *not fun.*

We don't need to transform the universe into something we feel *dutifully obligated* to create, but isn't really much fun - in the same way that a Christian would feel dutifully obliged to enjoy heaven - or that some strange folk think that creating orgasmium is, logically, the rightest thing to do.

Fun is okay. It's allowed. It doesn't get any better than fun.

And then we can turn our attention to the question of what *is* fun, and how to have it.

High Challenge

There's a class of prophecy that runs: "In the Future, machines will do all the work. Everything will be automated. Even labor of the sort we now consider 'intellectual', like engineering, will be done by machines. We can sit back and own the capital. You'll never have to lift a finger, ever again."

But then won't people be bored?

No; they can play computer games - not like *our* games, of course, but much more advanced and entertaining.

Yet wait! If you buy a modern computer game, you'll find that it contains some tasks that are - there's no kind word for this - *effortful.* (I would even say "difficult", with the understanding that we're talking about something that takes 10 minutes, not 10 years.)

So in the future, we'll have programs that *help*you play the game - taking over if you get stuck on the game, or just bored; or so that you can play games that would otherwise be too advanced for you.

But isn't there some wasted effort, here? Why have one programmer working to make the game harder, and another programmer to working to make the game easier? Why not just make the game easier to *start with?* Since you play the game to get gold and experience points, making the game easier will let you get more gold per unit time: the game will become more fun.

So this is the ultimate end of the prophecy of technological progress - just staring at a screen that says "YOU WIN", forever.

And maybe we'll build a robot that does *that,* too.

Then what?

The world of machines that do *all* the work - well, I don't want to say it's "analogous to the Christian Heaven" because it isn't supernatural; it's something that could in principle be realized. Religious analogies are far too easily tossed around as accusations... But, without implying any other similarities, I'll say that it seems analogous in the sense that eternal laziness "sounds like good news" to your present self who still has to work.

And as for playing games, as a substitute - what *is* a computer game except synthetic work? Isn't there a wasted step here? (And computer games in their present form, considered as work, have various aspects that reduce stress and increase engagement; but they also carry costs in the form of artificiality and isolation.)

I sometimes think that futuristic ideals phrased in terms of "getting rid of work" would be better reformulated as "removing low-quality work to make way for high-quality work".

There's a broad class of goals that aren't suitable as the long-term meaning of life, because you can actually achieve them, and then you're done.

To look at it another way, if we're looking for a suitable long-run meaning of life, we should look for goals that are good to *pursue* and not just good to *satisfy.*

Or to phrase that somewhat less paradoxically: We should look for valuations that are over 4D states, rather than 3D states. Valuable ongoing processes, rather than "make the universe have property P and then you're done".

Timothy Ferris is again worth quoting: To find happiness, "the question you should be asking isn't 'What do I want?' or 'What are my goals?' but 'What would excite me?'"

You might say that for a long-run meaning of life, we need games that are fun to *play* and not just to *win.*

Mind you - sometimes you *do* want to win. There are legitimate goals where winning is everything. If you're talking, say, about curing cancer, then the suffering experienced by even a single cancer patient outweighs any fun that you might have in solving their problems. If you work at creating a cancer cure for twenty years through your own efforts, learning new knowledge and new skill, making friends and allies - and then some alien superintelligence offers you a cancer cure on a silver platter for thirty bucks - then you shut up and take it.

But "curing cancer" is a problem of the 3D-predicate sort: you want the no-cancer predicate to go from False in the present to True in the future. The importance of this destination far outweighs the journey; you don't want to *go* there, you just want to *be* there. There are many *legitimate* goals of this sort, but they are not suitable as long-run fun. "Cure cancer!" is a worthwhile

activity for us to pursue here and now, but it is not a plausible future goal of galactic civilizations.

Why should this "valuable ongoing process" be a process of *trying to do things* - why not a process of passive experiencing, like the Buddhist Heaven?

I confess I'm not entirely sure how to set up a "passively experiencing" mind. The human brain was *designed* to perform various sorts of internal work that add up to an active intelligence; even if you lie down on your bed and exert no particular effort to think, the thoughts that go on through your mind are activities of brain areas that are designed to, you know, *solve problems.*

How much of the human brain could you eliminate, *apart*from the pleasure centers, and still keep the subjective experience of pleasure?

I'm not going to touch that one. I'll stick with the much simpler answer of "I wouldn't actually *prefer*to be a passive experiencer." If I *wanted* Nirvana, I might try to figure out how to achieve that impossibility. But once you strip away Buddha telling me that Nirvana is the end-all of existence, Nirvana seems rather more like "sounds like good news in the moment of first being told" or "ideological belief in desire" rather than, y'know, something I'd actually *want.**

The reason I have a mind at all, is that natural selection built me to *do* things - to solve certain kinds of problems.

"Because it's human nature" is not an explicit justification for anything. There is human nature, which is what we are; and there is humane nature, which is what, being human, we wish we were.

But I don't *want* to change my nature toward a more passive object - which *is*a justification. A happy blob is *not*what, being human, I wish to become.

I earlier argued that many values require both subjective happiness and the external objects of that happiness. That you can legitimately have a utility function that says, "It matters to me whether or not the person I love is a real human being or just a highly realistic nonsentient chatbot, *even if I don't know,* because that-which-I-value is not my own state of mind, but the external reality." So that you need both the experience of love, and the real lover.

You can similarly have valuable activities that require both real challenge and real effort.

Racing along a track, it matters that the other racers are real, and that you have a real chance to win or lose. (We're not talking about physical determinism here, but whether some external optimization process explicitly chose for you to win the race.)

And it matters that you're racing with your own skill at running and your own willpower, not just pressing a button that says "Win". (Though, since you never designed your own leg muscles, you *are* racing using strength that isn't

yours. A race between robot cars is a purer contest of their designers. There is plenty of room to improve on the human condition.)

And it matters that you, a sentient being, are experiencing it. (Rather than some nonsentient process carrying out a skeleton imitation of the race, trillions of times per second.)

There must be the true effort, the true victory, and the true experience - the journey, the destination and the traveler.

Complex Novelty

****From Greg Egan's *Permutation City*:

The workshop abutted a warehouse full of table legs - one hundred and sixty-two thousand, three hundred and twenty-nine, so far. Peer could imagine nothing more satisfying than reaching the two hundred thousand mark - although he knew it was likely that he'd change his mind and abandon the workshop before that happened; new vocations were imposed by his exoself at random intervals, but statistically, the next one was overdue. Immediately before taking up woodwork, he'd passionately devoured all the higher mathematics texts in the central library, run all the tutorial software, and then personally contributed several important new results to group theory - untroubled by the fact that none of the Elysian mathematicians would ever be aware of his work. Before that, he'd written over three hundred comic operas, with librettos in Italian, French and English - and staged most of them, with puppet performers and audience. Before that, he'd patiently studied the structure and biochemistry of the human brain for sixty-seven years; towards the end he had fully grasped, to his own satisfaction, the nature of the process of consciousness. Every one of these pursuits had been utterly engrossing, and satisfying, at the time. He'd even been interested in the Elysians, once.
No longer. He preferred to think about table legs.

Among science fiction authors, Greg Egan is my favorite; of Greg Egan's books, *Permutation City* is my favorite; and this particular passage in *Permutation City,* more than any of the others, I find utterly horrifying.

If this were all the hope the future held, I don't know if I could bring myself to try. Small wonder that people don't sign up for cryonics, if even SF writers think this is the best we can do.

You could think of this whole series on Fun Theory as my reply to Greg Egan - a list of the ways that his human-level uploaded civilizations Fail At Fun. (And yes, this series will also explain what's wrong with the Culture and how to fix it.)

We won't get to *all* of Peer's problems today - but really. *Table legs?*

I could see myself carving *one* table leg, maybe, if there was something non-obvious to learn from the experience. But not 162,329.

In *Permutation City*, Peer modified himself to find table-leg-carving fascinating and worthwhile and pleasurable. But really, at *that*point, you might as well modify yourself to get pleasure from playing Tic-Tac-Toe, or lie motionless on a pillow as a limbless eyeless blob having fantastic orgasms. It's not a worthy use of a human-level intelligence.

Worse, carving the 162,329th table leg doesn't *teach*you anything that you didn't already know from carving 162,328 previous table legs. A mind that changes so little in life's course is scarcely experiencing time.

But apparently, once you do a little group theory, write a few operas, and solve the mystery of consciousness, there isn't much else worth doing in life: you've *exhausted the entirety of Fun Space* down to the level of table legs.

Is this plausible? How large is Fun Space?

Let's say you were a human-level intelligence who'd never seen a Rubik's Cube, or anything remotely like it. As Hofstadter describes in two whole chapters of *Metamagical Themas,* there's a *lot* that intelligent human novices can learn from the Cube - like the whole notion of an "operator" or "macro", a sequence of moves that accomplishes a limited swap with few side effects. Parity, search, impossibility -

So you learn these things in the long, difficult course of solving the *first*scrambled Rubik's Cube you encounter. The *second*scrambled Cube - solving it might still be difficult, still be enough fun to be worth doing. But you won't have quite the same pleasurable shock of encountering something as new, and strange, and interesting as the first Cube was unto you.

Even if you encounter a variant of the Rubik's Cube - like a 4x4x4 Cube instead of a 3x3x3 Cube - or even a Rubik's Tesseract (a 3x3x3x3 Cube in four dimensions) - it still won't contain quite as much fun as the first Cube you ever saw. I haven't tried mastering the Rubik's Tesseract myself, so I don't know if there are added secrets in four dimensions - but it doesn't seem likely to teach me anything as fundamental as "operators", "side effects", or "parity".

(I was quite young when I encountered a Rubik's Cube in a toy cache, and so that actually *is* where I discovered such concepts. I tried that Cube on and off for months, without solving it. Finally I took out a book from the library on Cubes, applied the macros there, and discovered that this particular Cube was *unsolvable*- it had been disassembled and reassembled into an impossible position. I think I was faintly annoyed.)

Learning is fun, but it *uses up* fun: you can't have the same stroke of genius twice. Insight is insight because it makes future problems *less difficult,* and "deep" because it applies to many such problems.

And the smarter you are, the faster you learn - so the smarter you are, the less *total*fun you can have. Chimpanzees can occupy themselves for a lifetime at tasks that would bore you or I to tears. Clearly, the solution to Peer's difficulty is to become stupid enough that carving table legs is *difficult* again - and so lousy at generalizing that every table leg is a new and exciting challenge -

Well, but hold on: If you're a chimpanzee, you can't understand the Rubik's Cube *at all.* At least I'm willing to bet against anyone training a chimpanzee to solve one - let alone a chimpanzee solving it spontaneously - let alone a chimpanzee understanding the deep concepts like "operators", "side effects", and "parity".

I could be wrong here, but it seems to me, on the whole, that when you look at the number of ways that chimpanzees have fun, and the number of ways that humans have fun, *that Human Fun Space is larger than Chimpanzee Fun Space.*

And not in a way that increases just *linearly*with brain size, either.

The space of problems that are Fun to a given brain, *will*definitely be smaller than the exponentially increasing space of all possible problems that brain can *represent.* We are interested only in the borderland between triviality and impossibility - problems difficult enough to worthily occupy our minds, yet tractable enough to be worth challenging. (What *looks* "impossible" is not always impossible, but the border is still *somewhere* even if we can't see it at a glance - there are some problems so difficult you can't even learn much from failing.)

An even stronger constraint is that if you do something many times, you ought to learn from the experience and get better - many problems of the same *difficulty*will have the same "learnable lessons" embedded in them, so that doing one consumes some of the fun of others.

As you learn new things, and your skills improve, problems will get easier. Some will move off the border of the possible and the impossible, and become too easy to be interesting.

But *others*will move from the territory of impossibility into the borderlands of mere extreme difficulty. It's easier to invent group theory if you've solved the Rubik's Cube first. There are insights you can't have without prerequisite insights.

If you get smarter over time (larger brains, improved mind designs) that's a still higher octave of the same phenomenon. (As best I can grasp the Law, there are insights you can't understand *at all*without having a brain of sufficient size and sufficient design. Humans are not maximal in this sense, and I don't think there should be any maximum - but that's a rather deep topic, which I shall not explore further in this blog post. Note that Greg Egan seems to explicitly believe the reverse - that humans can understand *anything understandable* - which explains a lot.)

One suspects that in a better-designed existence, the eudaimonic rate of intelligence increase would be bounded below by the need to *integrate* the loot of your adventures - to incorporate new knowledge and new skills *efficiently*, without swamping your mind in a sea of disconnected memories and associations - to manipulate larger, more powerful concepts that generalize more of your accumulated life-knowledge at once.

And one also suspects that part of the poignancy of transhuman existence will be having to *move on* from your current level - get smarter, leaving old challenges behind - before you've explored more than an infinitesimal fraction of the Fun Space for a mind of your level. If, like me, you play through computer games trying to slay every single monster so you can collect every single experience point, this is as much tragedy as an improved existence could possibly need.

Fun Space can increase much more slowly than the space of representable problems, and still overwhelmingly swamp the amount of time you could bear to spend as a mind of a fixed level. Even if Fun Space grows at some ridiculously tiny rate like N-squared - bearing in mind that the actual raw space of representable problems goes as 2^N - we're still talking about "way more fun than you can handle".

If you consider the loot of every human adventure - everything that was ever learned about science, and everything that was ever learned about people, and all the original stories ever told, and all the original games ever invented, and all the plots and conspiracies that were ever launched, and all the personal relationships ever raveled, and all the ways of existing that were ever tried, and all the glorious epiphanies of wisdom that were ever minted -

- and you deleted all the duplicates, keeping only one of every lesson that had the same moral -
- how long would you have to stay human, to collect *every* gold coin in the dungeons of history?

Would it all fit into a single human brain, without that mind completely disintegrating under the weight of unrelated associations? And even then, would you have come close to exhausting the space of *human possibility,*which we've surely not finished exploring?

This is all sounding like suspiciously good news. So let's turn it around. Is there any way that Fun Space could fail to grow, and instead collapse?

Suppose there's only so many deep insights you *can* have on the order of "parity", and that you collect them all, and then math is never again as exciting as it was in the beginning. And that you then exhaust the shallower insights, and the trivial insights, until finally you're left with the delightful shock of "Gosh wowie gee willickers, the product of 845 and 109 is 92105, I didn't know that logical truth before."

Well - obviously, if you sit around and catalogue all the deep insights *known* to you to exist, you're going to end up with a bounded list. And equally obviously, if you declared, "This is all there is, and all that will ever be," you'd be taking an unjustified step. (Though I fully expect some people out there to step up and say how it seems to them that they've already started to run out of available insights that are as deep as the ones they remember from their childhood. And I fully expect that - compared to the sort of person who makes such a pronouncement - I *personally* will have collected more additional insights than they believe exist in the whole remaining realm of possibility.)

Can we say anything more on this subject of fun insights that might exist, but that we haven't yet found?

The obvious thing to do is start appealing to Godel, but Godelian arguments are dangerous tools to employ in debate. It does seem to me that Godelian arguments weigh in the general direction of "inexhaustible deep insights", but inconclusively and only by loose analogies.

For example, the Busy-Beaver(N) problem asks for the longest running time of a Turing machine with no more than N states. The Busy Beaver problem is uncomputable - there is no fixed Turing machine that computes it for all N - because if you knew all the Busy Beaver numbers, you would have an infallible way of telling whether a Turing machine halts; just run it up for as long as the longest-running Turing machine of that size.

The human species has managed to figure out and prove the Busy Beaver numbers up to 4, and they are:

BB(1): 1
BB(2): 6
BB(3): 21
BB(4): 107

Busy-Beaver 5 is believed to be 47,176,870.

The current lower bound on Busy-Beaver(6) is $\sim 2.5 \times 10^{2879}$.

This function *provably* grows faster than any compact specification you can imagine. Which would seem to argue that each new Turing machine is exhibiting a new and interesting kind of behavior. Given infinite time, you would even be able to *notice* this behavior. You won't ever know for certain that you've discovered the Busy-Beaver *champion* for any given N, after finite time; but conversely, you will *notice* the Busy Beaver champion for any N after some finite time.

Yes, this is an *unimaginably long* time - one of the few occasions where the word "unimaginable" is literally correct. We can't *actually* do this unless reality works the way it does in Greg Egan novels. But the point is that in the limit of infinite time we can point to *something* sorta like "an infinite sequence of *learnable* deep insights not reducible to any of their predecessors or to any learnable abstract summary". It's not conclusive, but it's at least *suggestive.*

Now you could still look at that and say, "I don't think my life would be an adventure of neverending excitement if I spent until the end of time trying to figure out the weird behaviors of slightly larger Tuing machines."

Well - as I said before, Peer is doing more than *one*thing wrong. Here I've dealt with only one sort of dimension of Fun Space - the dimension of how much *novelty* we can expect to find available to introduce into our fun.

But even on the arguments given so far... I don't call it conclusive, but it seems like sufficient reason to *hope and expect* that our descendants and future selves won't exhaust Fun Space to the point that there is literally nothing left to do but carve the 162,329th table leg.

Continuous Improvement

When is it adaptive for an organism to be satisfied with what it has? When does an organism have enough children *and* enough food? The answer to the second question, at least, is obviously "never" from an evolutionary standpoint. The first proposition might be true if the reproductive risks of all available options exceed their reproductive benefits. In general, though, it is a rare organism in a rare environment whose reproductively optimal strategy is to rest with a smile on its face, feeling happy.

To a first approximation, we might say something like "The evolutionary purpose of emotion is to direct the cognitive processing of the organism toward achievable, reproductively relevant goals". *Achievable*goals are usually located in the Future, since you can't affect the Past. Memory is a useful trick, but *learning the lesson* of a success or failure isn't the same goal as the original event - and usually the emotions associated with the memory are less intense than those of the original event.

Then the way organisms and brains are built right now, "true happiness" might be a chimera, a carrot dangled in front of us to make us take the next step, and then yanked out of our reach as soon as we achieve our goals.

This hypothesis is known as the hedonic treadmill.

The famous pilot studies in this domain demonstrated e.g. that past lottery winners' stated subjective well-being was not significantly greater than that of an average person, after a few years or even months. Conversely, accident victims with severed spinal cords were not as happy as before the accident after six months - around 0.75 sd less than control groups - but they'd still adjusted much more than they had expected to adjust.

This being the transhumanist form of Fun Theory, you might perhaps say: "Let's get rid of this effect. Just delete the treadmill, at least for positive events."

I'm not *entirely* sure we can get away with this. There's the possibility that comparing good events to not-as-good events is what gives them part of their

subjective quality. And on a moral level, it sounds perilously close to tampering with Boredom itself.

So suppose that instead of modifying minds and values, we first ask what we can do by modifying the environment. Is there enough fun in the universe, sufficiently accessible, for a transhuman to *jog off the hedonic treadmill* - improve their life *continuously*, at a sufficient rate to leap to an even higher hedonic level before they had a chance to get bored with the previous one?

This question leads us into great and interesting difficulties.

I had a nice vivid example I wanted to use for this, but unfortunately I couldn't find the exact numbers I needed to illustrate it. I'd wanted to find a figure for the total mass of the neurotransmitters released in the pleasure centers during an average male or female orgasm, and a figure for the density of those neurotransmitters - density in the sense of mass/volume of the chemicals themselves. From this I could've calculated *how long a period of exponential improvement* would be possible - how many years you could have "the best orgasm of your life" by a margin of at least 10%, at least once per year - before your orgasm collapsed into a black hole, the total mass having exceeded the mass of a black hole with the density of the neurotransmitters.

Plugging in some random/Fermi numbers instead:

Assume that a microgram of additional neurotransmitters are released in the pleasure centers during a standard human orgasm. And assume that neurotransmitters have the same density as water. Then an orgasm can reach around 10^8 solar masses before it collapses and forms a black hole, corresponding to 10^{47} baseline orgasms. If we assume that a 100mg dose of crack is as pleasurable as 10 standard orgasms, then the street value of your last orgasm is around a hundred billion trillion trillion trillion dollars.

I'm sorry. I just had to do that calculation.

Anyway... requiring an *exponential* improvement eats up a factor of 10^{47} in short order. Starting from human standard and improving at 10% per year, it would take less than 1,200 years.

Of course you say, "This but shows the folly of brains that use an analog representation of pleasure. Go digital, young man!"

If you redesigned the brain to represent the intensity of pleasure using IEEE 754 double-precision floating-point numbers, a mere 64 bits would suffice to feel pleasures up to 10^308 hedons... in, um, whatever base you were using.

This still represents less than 7500 years of 10% annual improvement from a 1-hedon baseline, but after that amount of time, you can switch to larger floats.

Now we *have* lost a bit of fine-tuning by switching to IEEE-standard hedonics. The 64-bit double-precision float has an 11-bit exponent and a 52-bit fractional part (and a 1-bit sign). So we'll only have 52 bits of precision (16

decimal places) with which to represent our pleasures, however great they may be. An original human's orgasm would soon be lost in the rounding error... which raises the question of how we can *experience* these invisible hedons, when the finite-precision bits are the whole substance of the pleasure.

We also have the odd situation that, starting from 1 hedon, flipping a single bit in your brain can make your life 10^{154} times more happy.

And Hell forbid you flip the sign bit. Talk about a need for cosmic ray shielding.

But really - if you're going to go so far as to use imprecise floating-point numbers to represent pleasure, why stop there? Why not move to Knuth's up-arrow notation?

For that matter, IEEE 754 provides special representations for +/- INF, that is to say, positive and negative infinity. What happens if a bit flip makes you experience infinite pleasure? Does that mean you Win The Game?

Now all of these questions I'm asking are in some sense unfair, because right now I don't know exactly what I have to do with *any*structure of bits in order to turn it into a "subjective experience". Not that this is the right way to phrase the question. It's not like there's a ritual that summons some incredible density of positive qualia that could collapse in its own right and form an epiphenomenal black hole.

But don't laugh - or at least, don't *only* laugh - because in the long run, these are *extremely*important questions.

To give you some idea of what's at stake here, Robin, in "For Discount Rates", pointed out that an investment earning 2% annual interest for 12,000 years adds up to a googol (10^100) times as much wealth; therefore, "very distant future times are ridiculously easy to help via investment".

I observed that there weren't a googol atoms in the observable universe, let alone within a 12,000-lightyear radius of Earth.

And Robin replied, "I know of no law limiting economic value per atom."

If you've got an increasingly large number of bits - things that can be one or zero - and you're doing a proportional number of computations with them... then how fast can you grow the amount of fun, or pleasure, or value?

This echoes back to the questions in Complex Novelty, which asked how many kinds of problems and novel solutions you could find, and how many deep insights there were to be had. I argued there that the growth rate is *faster than linear* in bits, e.g., humans can have much more than four times as much fun as chimpanzees even though our absolute brain volume is only around four times theirs. But I don't think the growth in "depth of good insights" or "number of unique novel problems" is, um, *faster than exponential* in the size of the pattern.

Now... it might be that the Law simply permits outright that we can create very large amounts of subjective pleasure, every bit as substantial as the sort

of subjective pleasure we get now, by the expedient of writing down very large numbers in a digital pleasure center. In this case, we have got it made. Have we *ever* got it made.

In one sense I can definitely see where Robin is coming from. Suppose that you had a specification of the first 10,000 Busy Beaver machines - the longest-running Turing machines with 1, 2, 3, 4, 5... states. This list could easily fit on a small flash memory card, made up of a few measly avogadros of atoms.

And that small flash memory card would be worth...

Well, let me put it this way: If a mathematician said to me that the value of this memory card, was worth more than the rest of the entire observable universe minus the card... I wouldn't necessarily *agree with him outright.* But I would understand his point of view.

Still, I don't know if you can truly grok the fun contained in that memory card, without an unbounded amount of computing power with which to understand it. Ultradense information does not give you ultradense economic value or ultradense fun unless you can also *use* that information in a way that consumes few resources. Otherwise it's just More Fun Than You Can Handle.

Weber's Law of Just Noticeable Difference says that stimuli with an intensity scale, typically require a fixed fraction of proportional difference, rather than any fixed interval of absolute intensity, in order for the difference to be noticeable to a human or other organism. In other words, we may demand exponential increases because our *imprecise* brains can't *notice* smaller differences. This would suggest that our existing pleasures might already in effect possess a floating-point representation, with an exponent and a fraction - the army of actual neurons being used only to transmit an analog signal most of whose precision is lost. So we might be able to get away with using floats, even if we can't get away with using up-arrows.

But suppose that the inscrutable rules governing the substantiality of "subjective" pleasure actually require one neuron per hedon, or something like that.

Or suppose that we only choose to reward ourselves when we find a *better solution,* and that we don't choose to game the betterness metrics.

And suppose that we don't discard the Weber-Fechner law of "just noticeable difference", but go on demanding *percentage* annual improvements, year after year.

Or you might need to improve at a fractional rate in order to assimilate your own memories. Larger brains would lay down larger memories, and hence need to grow exponentially - efficiency improvements suiting to *moderate*the growth, but not to eliminate the exponent.

If fun or intelligence or value can only grow as fast as the mere *cube*of the brain size... and yet we demand a 2% improvement every year...

Then 350 years will pass before our resource consumption grows a single order of magnitude.

And yet there are only around 10^{80} atoms in the observable universe.

Do the math.

(It works out to a lifespan of around 28,000 years.)

Now... before everyone gets all depressed about this...

We can still hold out a fraction of hope for *real immortality*, aka "emortality". As Greg Egan put it, "Not dying after a very long time. Just not dying, period."

The laws of physics as we know them prohibit emortality on multiple grounds. It is a fair historical observation that, over the course of previous centuries, civilizations have become able to do things that previous civilizations called "physically impossible". This reflects a change in knowledge about the laws of physics, not a change in the actual laws; and we cannot do *everything* once thought to be impossible. We violate Newton's version of gravitation, but not conservation of energy. It's a good historical bet that the future will be able to do at least *one* thing our physicists would call impossible. But you can't bank on being able to violate any *particular* "law of physics" in the future.

There is just... a shred of reasonable hope, that our physics might be *much* more incomplete than we realize, or that we are wrong in exactly the right way, or that anthropic points I don't understand might come to our rescue and let us escape these physics (also *a la* Greg Egan).

So I haven't lost hope. But I haven't lost despair, either; *that*would be faith.

In the case where our resources really are limited and there is no way around it...

...the question of how fast a rate of *continuous improvement*you demand for an acceptable quality of life - an annual percentage increase, or a fixed added amount - and the question of *how much* improvement you can pack into patterns of linearly increasing size - adding up to the fun-theoretic question of how fast you have to expand your resource usage over time to lead a life worth living...

...determines the maximum lifespan of sentient beings.

If you can get by with increasing the size *in bits* of your mind at a linear rate, then you can last for quite a while. Until the end of the universe, in many versions of cosmology. *And* you can have a child (or two parents can have two children), and the children can have children. Linear brain size growth * linear population growth = quadratic growth, and cubic growth at lightspeed should be physically permissible.

But if you have to grow exponentially, in order for your ever-larger mind and its ever-larger memories not to end up uncomfortably squashed into too small

a brain - squashed down to a point, to the point of it being pointless - then a transhuman's life is measured in subjective eons at best, and more likely subjective millennia. Though it would be a merry life indeed.

My own eye has trouble enough looking ahead a mere century or two of growth. It's not like I can imagine any sort of *me* the size of a galaxy. I just want to live one more day, and tomorrow I will still want to live one more day. The part about "wanting to live forever" is just an induction on the positive integers, not an instantaneous vision whose desire spans eternity.

If I can see to the fulfillment of all my present self's goals that I can concretely envision, shouldn't that be enough for me? And my century-older self will also be able to see that far ahead. And so on through thousands of generations of selfhood until some distant figure the size of a galaxy has to depart the physics we know, one way or the other... Should that be *scary?*

Yeah, I hope like hell that emortality is possible.

Failing that, I'd at least like to find out one way or the other, so I can get on with my life instead of having that lingering uncertainty.

For now, one of the reasons I care about people alive today is the thought that if creating new people just divides up a finite pool of resource available *here*, but we live in a Big World where there are plenty of people *elsewhere*with their own resources... then we might not want to create so many new people *here*. Six billion now, six trillion at the end of time? Though this is more an idiom of linear growth than exponential - with exponential growth, a factor of 10 fewer people just buys you another 350 years of lifespan per person, or whatever.

But I do hope for emortality. Odd, isn't it? How abstract should a hope or fear have to be, before a human can stop thinking about it?

Oh, and finally - there's an idea in the literature of hedonic psychology called the "hedonic set point", based on identical twin studies showing that identical twins raised apart have highly similar happiness levels, more so than fraternal twins raised together, people in similar life circumstances, etcetera. There are things that do seem to shift your set point, but not much (and permanent downward shift happens more easily than permanent upward shift, what a surprise). Some studies have suggested that up to 80% of the variance in happiness is due to genes, or something shared between identical twins in different environments at any rate.

If *no*environmental improvement ever has much effect on subjective well-being, the way you are now, because you've got a more or less genetically set level of happiness that you drift back to, then...

Well, my usual heuristic is to imagine messing with environments before I imagine messing with minds.

But in this case? Screw that. That's just *stupid.* Delete it without a qualm.

Sensual Experience

Modern day gamemakers are constantly working on higher-resolution, more realistic graphics; more immersive sounds - but they're a long *long*way off real life.

Pressing the "W" key to run forward as a graphic of a hungry tiger bounds behind you, just doesn't seem quite as *sensual*as running frantically across the savanna with your own legs, breathing in huge gasps and pumping your arms as the sun beats down on your shoulders, the grass brushes your shins, and the air whips around you with the wind of your passage.

Don't mistake me for a luddite; I'm not saying the technology *can't* get that good. I'm saying it hasn't gotten that good *yet.*

Failing to escape the computer tiger would also have fewer *long-term consequences* than failing to escape a biological tiger - it would be less a part of the total story of your life - meaning you're also likely to be less *emotionally* involved. But that's a topic for another post. Today's post is just about the sensual quality of the experience.

Sensual experience isn't a question of some mysterious quality that only the "real world" possesses. A computer screen is as *real* as a tiger, after all. Whatever *is,* is real.

But the pattern of the pseudo-tiger, inside the computer chip, is nowhere near as complex as a biological tiger; it offers far fewer modes in which to interact. And the sensory bandwidth between you and the computer's pseudo-world is relatively low; and the information passing along it isn't in quite the right format.

It's not a question of computer tigers being "virtual" or "simulated", and therefore somehow a separate magisterium. But with present technology, and the way your brain is presently set up, you'd have *a lot more neurons* involved in running away from a biological tiger.

Running would fill your whole vision with motion, not just a flat rectangular screen - which translates into more square centimeters of visual cortex getting actively engaged.

The graphics on a computer monitor try to trigger your sense of spatial motion (residing in the parietal cortex, btw). But they're presenting the information differently from its native *format*- without binocular vision, for example, and without your vestibular senses indicating true motion. So the sense of *motion* isn't likely to be quite the same, what it would be if you were running.

And there's the sense of touch that indicates the wind on your skin; and the proprioceptive sensors that respond to the position of your limbs; and the nerves that record the strain on your muscles. There's a whole strip of sensorimotor cortex running along the top of your brain, that would be much more intensely involved in "real" running.

It's a very old observation, that *Homo sapiens* was made to hunt and gather on the savanna, rather than work in an office. Civilization and its discontents... But alienation needs a causal mechanism; it doesn't just happen by magic. Physics is physics, so it's not that one environment is less real than another. But our *brains*are more adapted to *interfacing*with jungles than computer code.

Writing a complicated computer program carries its own triumphs and failures, heights of exultation and pits of despair. But is it the same sort of *sensual* experience as, say, riding a motorcycle? I've never actually ridden a motorcycle, but I expect not.

I've experienced the *exhilaration* of getting a program right on the dozenth try after finally spotting the problem. I doubt a random moment of a motorcycle ride actually feels *better* than that. But still, my hunter-gatherer ancestors never wrote computer programs. And so my mind's grasp on code is maintained using more rarefied, more abstract, more general capabilities - which means less sensual involvement.

Doesn't computer programming *deserve* to be as much of a sensual experience as motorcycle riding? Some time ago, a relative once asked me if I thought that computer programming could use all my talents; I at once replied, "There is *no limit* to the talent you can use in computer programming." It's as close as human beings have ever come to playing with the raw stuff of creation - but our grasp on it is too distant from the jungle. All our involvement is through letters on a computer screen. I win, and I'm happy, but there's no wind on my face.

If only my ancestors back to the level of my last common ancestor with a mouse, had constantly faced the challenge of writing computer programs! Then I would have brain areas suited to the task, and programming computers would be more of a sensual experience...

Perhaps it's not too late to fix the mistake?

If there were something around that was smart enough to rewrite human brains without breaking them - not a trivial amount of smartness - then it would be possible to expand the range of things that are *sensually* fun.

Not just novel challenges, but novel high-bandwidth senses and corresponding new brain areas. Widening the sensorium to include new vivid, detailed experiences. And not neglecting the other half of the equation, high-bandwidth motor connections - new motor brain areas, to control with subtlety our new limbs (the parts of the process that we control as our direct handles on it).

There's a story - old now, but I remember how exciting it was when the news first came out - about a brain-computer interface for a "locked-in" patient (who could previously only move his eyes), connecting control of a computer cursor directly to neurons in his visual cortex. It took some training at first for him to use the cursor - he started out by trying to move his paralyzed arm, which was

the part of the motor cortex they were interfacing, and watched as the cursor jerked around on the screen. But after a while, they asked the patient, "What does it feel like?" and the patient replied, "It doesn't feel like anything." He just controlled the cursor the same sort of way he would have controlled a finger, except that it wasn't a finger, it was a cursor.

Like most brain modifications, adding new senses is not something to be done lightly. Sensual experience too *easily* renders a task involving.

Consider taste buds. Recognizing the taste of the same food on different occasions was very important to our ancestors - it was how they *learned*what to eat, that extracted regularity. And our ancestors also got helpful reinforcement from their taste buds about what to eat - reinforcement which is now worse than useless, because of the marketing incentive to reverse-engineer tastiness using artificial substances. By now, it's probably true that at least some people have eaten 162,329 potato chips in their lifetimes. That's even less novelty and challenge than carving 162,329 table legs.

I'm not saying we should try to eliminate our senses of taste. There's a lot to be said for grandfathering in the senses we started with - it preserves our existing life memories, for example. Once you realize how easy it would be for a mind to collapse into a pleasure center, you start to respect the "complications" of your goal system a lot more, and be more wary around "simplifications".

But I do want to nudge people into adopting something of a questioning attitude toward the senses we have now, rather than assuming that the existing senses are The Way Things Have Been And Will Always Be. A sex organ bears thousands of densely packed nerves for signal strength, but that signal - however strong - isn't as *complicated* as the sensations sent out by taste buds. Is that really appropriate for one of the most interesting parts of human existence? That even a novice chef can create a wider variety of taste sensations for your tongue, than - well, I'd better stop there. But from a fun-theoretic standpoint, the existing setup is wildly unbalanced in a lot of ways. It wasn't *designed* for the sake of eudaimonia.

I conclude with the following cautionary quote from an old IRC conversation, as a reminder that maybe not *everything*should be a sensual experience:

<MRAmes> I want a sensory modality for regular expressions.

Living By Your Own Strength

"Myself, and Morisato-san... we want to live together by our own strength."

Jared Diamond once called agriculture "the worst mistake in the history of the human race". Farmers could grow more wheat than hunter-gatherers could collect nuts, but the evidence seems pretty conclusive that agriculture traded quality of life for quantity of life. One study showed that the farmers in an area

were six inches shorter and seven years shorter-lived than their hunter-gatherer predecessors - even though the farmers were more numerous.

I don't know if I'd call agriculture a *mistake*. But one should at least be aware of the downsides. Policy debates should not appear one-sided.

In the same spirit -

Once upon a time, our hunter-gatherer ancestors strung their own bows, wove their own baskets, whittled their own flutes.

And part of our alienation from that environment of evolutionary adaptedness, is the number of tools we use that we don't understand and couldn't make for ourselves.

You can look back on *Overcoming Bias*, and see that I've always been suspicious of borrowed strength. (Even before I understood the source of Robin's and my disagreement about the Singularity, that is.) In Guessing the Teacher's Password I talked about the (well-known) problem in which schools end up teaching verbal behavior rather than real knowledge. In Truly Part of You I suggested one test for false knowledge: Imagine deleting a fact from your mind, and ask if it would grow back.

I know many ways to prove the Pythagorean Theorem, including at least one proof that is purely visual and can be seen at a glance. But if you deleted the Pythagorean Theorem from my mind *entirely,* would I have enough math skills left to grow it back the next time I needed it? I hope so - certainly I've solved math problems that *seem* tougher than that, what with benefit of hindsight and all. But, as I'm not an AI, I can't actually switch off the memories and associations, and test myself in that way.

Wielding someone else's strength to do things beyond your own understanding - that really *is* as dangerous as the Deeply Wise phrasing makes it sound.

I observed in Failing to Learn from History (musing on my childhood foolishness in offering a mysterious answer to a mysterious question): "If only I had *personally* postulated astrological mysteries and then discovered Newtonian mechanics, postulated alchemical mysteries and then discovered chemistry, postulated vitalistic mysteries and then discovered biology. I would have thought of my Mysterious Answer and said to myself: *No way am I falling for that again."*

At that point in my childhood, I'd been *handed* some techniques of rationality but I didn't exactly *own* them. Borrowing someone else's knowledge really doesn't give you anything *remotely*like the same power level required to discover that knowledge for yourself.

Would Isaac Newton have remained a mystic, even in that earlier era, if he'd *lived*the lives of Galileo and Archimedes instead of just reading about them? If he'd *personally*seen the planets reduced from gods to spheres in a telescope? If he'd *personally*fought that whole war against ignorance and mystery that had

to be fought, before Isaac Newton could be handed math and science as a start to his further work?

We stand on the shoulders of giants, and in doing so, the power that we wield is far out of proportion to the power that we could generate for ourselves. This is true even of our revolutionaries. And yes, we couldn't begin to support this world if people could only use their own strength. Even so, we are losing something.

That thought occurred to me, reading about the Manhattan Project, and the petition that the physicists signed to avoid dropping the atomic bomb on Japan. It was too late, of course; they'd already built the bombs and handed them to the military, and they couldn't take back that gift. And so nuclear weapons passed into the hands of politicians, who could never have created such a thing through their own strength...

Not that I'm saying the world would necessarily have been a better place, if physicists had possessed sole custody of ICBMs. What does a physicist know about international diplomacy, or war? And it's not as if Leo Szilard - who first thought of the fission chain reaction - had personally invented science; he too was using powers beyond his own strength. And it's not as if the physicists on the Manhattan Project raised the money to pay for their salaries and their materials; they were borrowing the strength of the politicians...

But if no one had been able to use nuclear weapons without, say, possessing the discipline of a scientist *and* the discipline of a politician - without personally knowing enough to construct an atomic bomb *and* make friends - the world might have been a slightly safer place.

And if nobody had been able to construct an atomic bomb without first discovering for themselves the nature and existence of physics, then we would have been *much*safer from atomic bombs, because no one would have been able to build them until they were two hundred years old.

With humans leaving the game after just seventy years, we couldn't support this world using only our own strengths. But we have traded quality of insight for quantity of insight.

It does sometimes seem to me that many of this world's problems, stem from our using powers that aren't appropriate to seventy-year-olds.

And there is a higher level of strength-ownership, which no human being has yet achieved. Even when we run, we're just using the muscles that evolution built for us. Even when we think, we're just using the brains that evolution built for us.

I'm not suggesting that people should create themselves from scratch without a starting point. Just pointing out that it would be a different world if we understood our own brains and could redesign our own legs. As *yet*there's no human "rationalist" or "scientist", whatever they know about "how to think",

who could actually *build a rational AI* - which shows you the limits of our self-understanding.

This is not the sort of thing that I'd suggest as an immediate alteration. I'm not suggesting that people should *instantly* on a silver platter be given full knowledge of how their own brains work and the ability to redesign their own legs. Because maybe people will be better off if they aren't *given*that kind of power, but rather have to work out the answer for *themselves.*

Just in terms of anomie versus fun, there's a big difference between being able to do things for yourself, and having to rely on other people to do them for you. (Even if you're doing them with a brain you never designed yourself.)

I don't know if it's a principle that would stay until the end of time, to the children's children. Maybe better-designed minds could handle opaque tools without the anomie.

But it is part of the commonly retold prophecy of Artificial Intelligence and the Age of the Machine, that this era must be accompanied by *greater* reliance on things outside yourself, *more* incomprehensible tools into which you have *less*insight and *less* part in their creation.

Such a prophecy is not surprising. That *is*the way the trend has gone so far, in our culture that is too busy staying alive to optimize for fun. From the fire-starting tools that you built yourself, to the village candleseller, and then from the candleseller to the electric light that runs on strange mathematical principles and is powered by a distant generator... we are surrounded by things outside ourselves and strengths outside our understanding; we need them to stay alive, or we buy them because it's easier that way.

But with a sufficient surplus of power, you could start doing things the eudaimonic way. Start rethinking the life experience as a road to internalizing new strengths, instead of just trying to keep people alive efficiently.

A Friendly AI doesn't have to be a continuation of existing trends. It's not the Machine. It's not the alien force of technology. It's not mechanizing a factory. It's not a new gadget for sale. That's not where the shape comes from. What it *is*- is not easy to explain; but I'm reminded of doc Smith's description of the Lens as "the physical manifestation of a purely philosophical concept". That philosophical concept doesn't have to manifest as new buttons to press - if, on reflection, that's not what we would want.

Free to Optimize

Stare decisis is the legal principle which binds courts to follow precedent, retrace the footsteps of other judges' decisions. As someone previously condemned to an Orthodox Jewish education, where I gritted my teeth at the idea that

medieval rabbis would always be wiser than modern rabbis, I *completely missed* the rationale for *stare decisis*. I thought it was about respect for the past.

But shouldn't we presume that, in the presence of science, judges closer to the future will know more - have new facts at their fingertips - which enable them to make better decisions? Imagine if engineers respected the decisions of past engineers, not as a source of good suggestions, but as a binding precedent! - That was my original reaction. The standard rationale behind *stare decisis* came as a shock of revelation to me; it considerably increased my respect for the whole legal system.

This rationale is *jurisprudence constante:* The legal system must above all be *predictable,* so that people can execute contracts or choose behaviors knowing the legal implications.

Judges are not necessarily there to *optimize,* like an engineer. The purpose of law is not to make the world perfect. The law is there to provide a *predictable environment* in which people can optimize their *own*futures.

I was amazed at how a principle that at first glance seemed so completely Luddite, could have such an Enlightenment rationale. It was a "shock of creativity" - a solution that ranked high in my preference ordering and low in my search ordering, a solution that violated my previous surface generalizations. "Respect the past *just*because it's the past" would not have easily occurred to me as a good solution for *anything*.

There's a peer commentary in *Evolutionary Origins of Morality* which notes in passing that "other things being equal, organisms will choose to reward themselves over being rewarded by caretaking organisms". It's cited as the Premack principle, but the actual Premack principle looks to be something quite different, so I don't know if this is a bogus result, a misremembered citation, or a nonobvious derivation. If true, it's definitely interesting from a fun-theoretic perspective.

Optimization is the ability to squeeze the future into regions high in your preference ordering. Living by my own strength, means squeezing my own future - not perfectly, but still being able to grasp *some* of the relation between my actions and their consequences. This is the strength of a human.

If I'm being *helped,* then some other agent is also squeezing my future - optimizing me - in the same rough direction that I try to squeeze myself. This is "help".

A human helper is unlikely to steer every part of my future that I could have steered myself. They're not likely to have already exploited every connection between action and outcome that I can myself understand. They won't be able to squeeze the future *that tightly;* there will be slack left over, that I can squeeze for myself.

We have little experience with being "caretaken" across any substantial gap in intelligence; the closest thing that human experience provides us with is the

idiom of parents and children. Human parents are still human; they may be smart*er* than their children, but they can't predict the future or manipulate the kids in any fine-grained way.

Even so, it's an empirical observation that some human parents *do*help their children so much that their children don't become strong. It's not that there's *nothing* left for their children to do, but with a hundred million dollars in a trust fund, they don't *need* to do much - their remaining motivations aren't strong enough. Something like that depends on genes, not just environment - not every overhelped child shrivels - but conversely it depends on environment too, not just genes.

So, in considering the kind of "help" that can flow from relatively stronger agents to relatively weaker agents, we have two potential problems to track:

1. Help so strong that it optimizes away the links between the desirable outcome and your own choices.
2. Help that is *believed*to be so reliable, that it takes off the psychological pressure to use your own strength.

Since (2) revolves around *belief*, could you just lie about how reliable the help was? Pretend that you're not going to help when things get bad - but then if things do get bad, you help anyway? That trick didn't work too well for Alan Greenspan and Ben Bernanke.

A superintelligence might be able to pull off a better deception. But in terms of moral theory and eudaimonia - we *are*allowed to have preferences over external states of affairs, not just psychological states. This applies to "I want to *really* steer my own life, not just believe that I do", just as it applies to "I want to have a love affair with a fellow sentient, not just a puppet that I am deceived into thinking sentient". So if we can state firmly from a value standpoint that we don't want to be fooled this way, then *building*an agent which respects that preference is a mere matter of Friendly AI.

Modify people so that they *don't* relax when they believe they'll be helped? I usually try to think of how to modify environments before I imagine modifying any people. It's not that I want to stay the same person forever; but the issues are rather more fraught, and one might wish to take it slowly, at some eudaimonic rate of personal improvement.

(1), though, is the most interesting issue from a philosophicalish standpoint. It impinges on the confusion named "free will". Of which I have already untangled; see the posts referenced at top, if you're recently joining *OB*.

Let's say that I'm an ultrapowerful AI, and I use my knowledge of your mind and your environment to forecast that, if left to your own devices, you will make $999,750. But this does not satisfice me; it so happens that I want you to make

at least $1,000,000. So I hand you $250, and then you go on to make $999,750 as you ordinarily would have.

How much of your own strength have you just lived by?

The first view would say, "I made 99.975% of the money; the AI only helped 0.025% worth."

The second view would say, "Suppose I had entirely slacked off and done nothing. Then the AI would have handed me $1,000,000. So my attempt to *steer my own future* was an illusion; my future was already determined to contain $1,000,000."

Someone might reply, "Physics is deterministic, so your future is already determined no matter what you or the AI does -"

But the second view interrupts and says, "No, you're not confusing me that easily. I am within physics, so in order for my future to be determined by me, it must be determined by physics. The Past does not reach around the Present and determine the Future before the Present gets a chance - that is mixing up a timeful view with a timeless one. But if there's an AI that really *does* look over the alternatives before I do, and really *does* choose the outcome before I get a chance, then I'm really *not* steering my own future. The future is no longer *counterfactually dependent* on my decisions."

At which point the first view butts in and says, "But of course the future is counterfactually dependent on your actions. The AI gives you $250 and then leaves. As a physical fact, if you didn't work hard, you would end up with only $250 instead of $1,000,000."

To which the second view replies, "I one-box on Newcomb's Problem, so my counterfactual reads 'if my decision were to not work hard, the AI would have given me $1,000,000 instead of $250'."

"So you're saying," says the first view, heavy with sarcasm, "that if the AI had wanted me to make at least $1,000,000 and it had ensured this through the general policy of handing me $1,000,000 flat on a silver platter, leaving me to earn $999,750 through my own actions, for a total of $1,999,750 - that this AI would have interfered *less* with my life than the one who just gave me $250."

The second view thinks for a second and says "Yeah, actually. Because then there's a stronger counterfactual dependency of the final outcome on your own decisions. Every dollar you earned was a real added dollar. The second AI helped you more, but it constrained your destiny less."

"But if the AI had done exactly the same thing, because it *wanted* me to make exactly $1,999,750 -"

The second view nods.

"That sounds a bit scary," the first view says, "for reasons which have nothing to do with the usual furious debates over Newcomb's Problem. You're making

your utility function path-dependent on the detailed cognition of the Friendly AI trying to help you! You'd be okay with it if the AI only *could* give you $250. You'd be okay if the AI had decided to give you $250 through a decision process that had *predicted the final outcome in less detail,* even though you acknowledge that in principle your decisions may already be highly deterministic. How is a poor Friendly AI supposed to help you, when your utility function is dependent, not just on the outcome, not just on the Friendly AI's actions, but dependent on *differences of the exact algorithm* the Friendly AI uses to arrive at *the same decision?* Isn't your whole rationale of one-boxing on Newcomb's Problem that you only care about what works?"

"Well, that's a good point," says the second view. "But sometimes we only care about what works, and yet sometimes we *do* care about the journey as well as the destination. If I was trying to cure cancer, I wouldn't care how I cured cancer, or whether I or the AI cured cancer, just so long as it ended up cured. This *isn't* that kind of problem. This is the problem of the eudaimonic journey - it's the reason I care in the first place whether I get a million dollars through my own efforts or by having an outside AI hand it to me on a silver platter. My utility function is not up for grabs. If I desire not to be optimized too hard by an outside agent, the agent needs to respect that preference even if it depends on the details of how the outside agent arrives at its decisions. Though it's also worth noting that decisions *are*produced by algorithms - if the AI *hadn't* been using the algorithm of doing just what it took to bring me up to $1,000,000, it probably *wouldn't* have handed me exactly $250."

The desire *not to be optimized too hard by an outside agent* is one of the structurally nontrivial aspects of human morality.

But I can think of *a* solution, which unless it contains some terrible flaw not obvious to me, sets a lower bound on the goodness of a solution: any alternative solution adopted, ought to be at least this good or better.

If there is anything in the world that resembles a god, people will try to pray to it. It's human nature to such an extent that people will pray even if there aren't any gods - so you can imagine what would happen if there were! But people don't pray to gravity to ignore their airplanes, because it is understood how gravity works, and it is understood that gravity doesn't adapt itself to the needs of individuals. Instead they understand gravity and try to turn it to their own purposes.

So one possible way of helping - which may or may not be the best way of helping - would be the gift of a world that works on *improved rules,* where the rules are stable and understandable enough that people can manipulate them and optimize their own futures together. A nicer place to live, but free of meddling gods beyond that. I have yet to think of a form of help that is less poisonous to human beings - but I am only human.

Added: Note that *modern* legal systems score a low Fail on this dimension - no single human mind can even *know* all the regulations any more, let alone

optimize for them. Maybe a professional lawyer who did nothing else could memorize all the regulations applicable to them personally, but I doubt it. As Albert Einstein observed, any fool can make things more complicated; what takes intelligence is moving in the opposite direction.

Harmful Options

Barry Schwartz's The Paradox of Choice - which I haven't read, though I've read some of the research behind it - talks about how offering people *more choices* can make them *less happy*.

A simple intuition says this shouldn't ought to happen to rational agents: If your current choice is X, and you're offered an alternative Y that's worse than X, and you know it, you can always just go on doing X. So a rational agent shouldn't do worse by having more options. The more available actions you have, the more powerful you become - that's how it should ought to work.

For example, if an ideal rational agent is initially *forced* to take only box B in Newcomb's Problem, and is then offered the *additional* choice of taking both boxes A and B, the rational agent shouldn't *regret having more options*. Such regret indicates that you're "fighting your own ritual of cognition" which helplessly selects the worse choice once it's offered you.

But this intuition only governs *extremely* idealized rationalists, or rationalists in extremely idealized situations. Bounded rationalists can easily do worse with strictly more options, because they burn computing operations to evaluate them. You could write an invincible chess program in one line of Python if its only legal move were the winning one.

Of course Schwartz and co. are not talking about anything so pure and innocent as the *computing cost* of having more choices.

If you're dealing, not with an ideal rationalist, not with a bounded rationalist, but with a *human being* -

Say, would you like to finish reading this post, or watch this surprising video instead?

Schwartz, I believe, talks primarily about the decrease in *happiness* and *satisfaction* that results from having more mutually exclusive options. Before this research was done, it was already known that people are more sensitive to losses than to gains, generally by a factor of between 2 and 2.5 (in various different experimental scenarios). That is, the pain of losing something is between 2 and 2.5 times as worse as the joy of gaining it. (This is an interesting constant in its own right, and may have something to do with compensating for our systematic overconfidence.)

So - if you can only choose one dessert, you're likely to be happier choosing from a menu of two than a menu of fourteen. In the first case, you eat one

dessert and pass up one dessert; in the latter case, you eat one dessert and pass up thirteen desserts. And we are more sensitive to loss than to gain.

(If I order dessert on a menu at all, I will order quickly and then close the menu and put it away, so as not to look at the other items.)

Not only that, but if the options have incommensurable attributes, then whatever option we select is likely to *look worse* because of the comparison. A luxury car that would have looked great by comparison to a Crown Victoria, instead becomes slower than the Ferrari, more expensive than the 9–5, with worse mileage than the Prius, and not looking quite as good as the Mustang. So we lose on satisfaction with the road we *did* take.

And then there are more direct forms of harm done by painful choices. IIRC, an experiment showed that people who *refused* to eat a cookie - who were offered the cookie, and chose *not* to take it - did worse on subsequent tests of mental performance than either those who ate the cookie or those who were not offered any cookie. You pay a price in mental energy for resisting temptation.

Or consider the various "trolley problems" of ethical philosophy - a trolley is bearing down on 5 people, but there's one person who's very fat and can be pushed onto the tracks to stop the trolley, that sort of thing. If you're forced to choose between two unacceptable evils, you'll pay a price either way. Vide *Sophie's Choice*.

An option need not be taken, or even be strongly considered, in order to wreak harm. Recall the point from "High Challenge", about how offering to do someone's work for them is not always helping them - how the ultimate computer game is not the one that just says "YOU WIN", forever.

Suppose your computer games, in addition to the *long difficult* path to your level's goal, also had little side-paths that you could use - directly in the game, as corridors - that would bypass all the enemies and take you straight to the goal, offering along the way all the items and experience that you could have gotten the hard way. And this corridor is always visible, out of the corner of your eye.

Evenif you resolutely refused to take the easy path through the game, knowing that it would cheat you of the very experience that you paid money in order to buy - wouldn't that always-visible corridor, make the game that much less fun? Knowing, for every alien you shot, and every decision you made, that there was always an easier path?

I don't know if this story has ever been written, but you can imagine a Devil who follows someone around, making their life miserable, *solely by offering them options which are never actually taken* - a "deal with the Devil" story that only requires the Devil to have the *capacity* to grant wishes, rather than ever granting a single one.

And what if the worse option is actually taken? I'm not suggesting that it is always a good idea for human governments to go around Prohibiting tempta-

tions. But the literature of heuristics and biases is replete with examples of reproducible stupid choices; and there is also such a thing as akrasia (weakness of will).

If you're an agent operating from a *much*higher vantage point - high enough to see humans as flawed algorithms, so that it's not a matter of second-guessing but second-*knowing* - then is it *benevolence* to offer choices that will assuredly be made wrongly? Clearly, removing all choices from someone and reducing their life to Progress Quest, is not helping them. But are we wise enough to *know when we should choose?* And in some cases, even offering that much of a choice, *even if the choice is made correctly,* may already do the harm...

Devil's Offers

An iota of fictional evidence from *The Golden Age* by John C. Wright:

Helion had leaned and said, "Son, once you go in there, the full powers and total command structures of the Rhadamanth Sophotech will be at your command. You will be invested with godlike powers; but you will still have the passions and distempers of a merely human spirit. There are two temptations which will threaten you. First, you will be tempted to remove your human weaknesses by abrupt mental surgery. The Invariants do this, and to a lesser degree, so do the White Manorials, abandoning humanity to escape from pain. Second, you will be tempted to indulge your human weakness. The Cacophiles do this, and to a lesser degree, so do the Black Manorials. Our society will gladly feed every sin and vice and impulse you might have; and then stand by helplessly and watch as you destroy yourself; because the first law of the Golden Oecumene is that no peaceful activity is forbidden. Free men may freely harm themselves, provided only that it is only themselves that they harm."
Phaethon knew what his sire was intimating, but he did not let himself feel irritated. Not today. Today was the day of his majority, his emancipation; today, he could forgive even Helion's incessant, nagging fears.
Phaethon also knew that most Rhadamanthines were not permitted to face the Noetic tests until they were octogenerians; most did not pass on their first attempt, or even their second. Many folk were not trusted with the full powers of an adult until they reached their Centennial. Helion, despite criticism from the other Silver-Gray branches, was permitting Phaethon to face the tests five years early...

Then Phaethon said, "It's a paradox, Father. I cannot be, at the same time and in the same sense, a child and an adult. And, if I am an adult, I cannot be, at the same time, free to make my own successes, but not free to make my own mistakes."
Helion looked sardonic. "'Mistake' is such a simple word. An adult who suffers a moment of foolishness or anger, one rash moment, has time enough to delete

or destroy his own free will, memory, or judgment. No one is allowed to force a cure on him. No one can restore his sanity against his will. And so we all stand quietly by, with folded hands and cold eyes, and meekly watch good men annihilate themselves. It is somewhat... quaint... to call such a horrifying disaster a 'mistake.'"

Is this the best Future we could possibly get to - the Future where you must be absolutely stern and resistant throughout your entire life, because *one moment of weakness* is enough to betray you to overwhelming temptation?

Such flawless perfection would be easy enough for a superintelligence, perhaps - for a *true*adult - but for a human, even a hundred-year-old human, it seems like a dangerous and inhospitable place to live. Even if you are strong enough to always choose correctly - maybe you don't want to *have*to be so strong, always at every moment.

This is the great flaw in Wright's otherwise shining Utopia - that the Sophotechs are *helpfully*offering up overwhelming temptations to people who would not be at *quite*so much risk from only *themselves*. (Though if not for this flaw in Wright's Utopia, he would have had no story...)

If I recall correctly, it was while reading *The Golden Age* that I generalized the principle "Offering people powers beyond their own is not always helping them."

If you couldn't just ask a Sophotech to edit your neural networks - and you couldn't buy a standard package at the supermarket - but, rather, had to study neuroscience yourself until you could do it with your own hands - then that would act as something of a natural limiter. Sure, there are pleasure centers that would be relatively easy to stimulate; but we don't tell you where they are, so you have to do your own neuroscience. Or we don't sell you your own neurosurgery kit, so you have to build it yourself - metaphorically speaking, anyway -

But you see the idea: it is not so terrible a disrespect for free will, to live in a world in which people are free to shoot their feet off *through their own strength* - in the hope that by the time they're smart enough to do it *under their own power*, they're smart enough *not*to.

The more dangerous and destructive the act, the more you require people to do it without external help. If it's really dangerous, you don't just require them to do their own engineering, but to do their own science. A singleton might be justified in prohibiting standardized textbooks in certain fields, so that people have to do their own science - make their own discoveries, learn to rule out their own stupid hypotheses, and fight their own overconfidence. Besides, everyone should experience the joy of major discovery at least once in their lifetime, and to do this properly, you may have to prevent spoilers from entering the public discourse. So you're getting three social benefits at once, here.

But now I'm trailing off into plots for SF novels, instead of Fun Theory per se. (It can be fun to muse how I would create the world if I had to order it

according to my own childish wisdom, but in real life one rather prefers to avoid that scenario.)

As a matter of Fun Theory, though, you can imagine a *better* world than the Golden Oecumene depicted above - it is not the *best* world imaginable, fun-theoretically speaking. We would prefer (if attainable) a world in which people own their own mistakes and their own successes, and yet they are not given loaded handguns on a silver platter, nor do they perish through suicide by genie bottle.

Once you imagine a world in which people can shoot off their own feet *through their own strength,* are you making that world incrementally better by offering incremental help along the way?

It's one matter to prohibit people from using dangerous powers that they have grown enough to acquire naturally - to literally *protect them from themselves.* One expects that if a mind kept getting smarter, at some eudaimonic rate of intelligence increase, then - if you took the most obvious course - the mind would eventually become able to edit its own source code, and bliss itself out if it chose to do so. Unless the mind's growth were steered onto a non-obvious course, or monitors were mandated to prohibit that event... To protect people *from their own powers* might take some twisting.

To descend from above and *offer dangerous powers as an untimely gift*, is another matter entirely. That's why the title of this post is "Devil's Offers", not "Dangerous Choices".

And to allow dangerous powers to be sold in a marketplace - or alternatively to prohibit them from being transferred from one mind to another - that is somewhere in between.

John C. Wright's writing has a particular poignancy for me, for in my foolish youth I thought that something very much like this scenario was a good idea - that a benevolent superintelligence ought to go around offering people lots of options, and doing as it was asked.

In retrospect, this was a case of a pernicious distortion where you end up believing things that are easy to market to other people.

I know someone who drives across the country on long trips, rather than flying. Air travel scares him. Statistics, naturally, show that flying a given distance is much safer than driving it. But some people fear too much the *loss of control* that comes from not having their own hands on the steering wheel. It's a common complaint.

The future sounds less scary if you imagine yourself having lots of control over it. For every awful thing that you imagine happening to you, you can imagine, "But I won't choose that, so it will be all right."

And if it's not your own hands on the steering wheel, you think of scary things, and imagine, "What if this is chosen *for* me, and I can't say no?"

But in real life rather than imagination, human choice is a fragile thing. If the whole field of heuristics and biases teaches us anything, it surely teaches us that. Nor has it been the verdict of experiment, that humans correctly estimate the flaws of their own decision mechanisms.

I flinched away from that thought's implications, not so much because I feared superintelligent paternalism *myself*, but because I feared what other people would say of that position. If I believed it, I would have to defend it, so I managed not to believe it. Instead I told people not to worry, a superintelligence would surely respect their decisions (and even believed it myself). A very pernicious sort of self-deception.

Human governments are made up of humans who are foolish like ourselves, plus they have poor incentives. Less skin in the game, and specific human brainware to be corrupted by wielding power. So we've learned the historical lesson to be wary of ceding control to human bureaucrats and politicians. We may even be emotionally hardwired to resent the loss of anything we perceive as power.

Which is just to say that people are biased, by instinct, by anthropomorphism, and by narrow experience, to *under*estimate how much they could potentially trust a superintelligence which lacks a human's corruption circuits, doesn't easily make certain kinds of mistakes, and has strong overlap between its motives and your own interests.

Do you trust yourself? Do you trust yourself to know when to trust yourself? If you're dealing with a superintelligence kindly enough to care about you at all, rather than disassembling you for raw materials, are you wise to second-guess its choice of *who*it thinks should decide? Do you think you have a superior epistemic vantage point here, or what?

Obviously we should not trust all agents who claim to be trustworthy - especially if they are *weak*enough, relative to us, to *need*our goodwill. But I am quite ready to accept that a benevolent superintelligence may not offer certain choices.

If you *feel safer* driving than flying, because that way it's your own hands on the steering wheel, statistics be damned -

- then maybe it isn't *helping*you, for a superintelligence to offer you the option of driving.

Gravity doesn't ask you if you would like to float up out of the atmosphere into space and die. But you don't go around complaining that gravity is a tyrant, right? You can build a spaceship if you work hard and study hard. It would be a more dangerous world if your six-year-old son could do it in an hour using string and cardboard.

Nonperson Predacates

There is a subproblem of Friendly AI which is so scary that I usually don't talk about it, because only a longtime reader of *Overcoming Bias* would react to it appropriately - that is, by saying, "Wow, that does sound like an *interesting*problem", instead of finding one of many subtle ways to scream and run away.

This is the problem that if you create an AI and tell it to model the world around it, it may form models of people that are people themselves. Not necessarily the *same* person, but people nonetheless.

If you look up at the night sky, and see the tiny dots of light that move over days and weeks - *planētoi*, the Greeks called them, "wanderers" - and you try to predict the movements of those planet-dots as best you can...

Historically, humans went through a journey as long and as wandering as the planets themselves, to find an accurate model. In the beginning, the models were things of cycles and epicycles, not much resembling the true Solar System.

But eventually we found laws of gravity, and finally built models - even if they were just on paper - that were *extremely* accurate so that Neptune could be deduced by looking at the unexplained perturbation of Uranus from its expected orbit. This required moment-by-moment modeling of where a simplified version of Uranus would be, and the other known planets. Simulation, not just abstraction. Prediction through simplified-yet-still-detailed pointwise similarity.

Suppose you have an AI that is around human beings. And like any Bayesian trying to explain its enivornment, the AI goes in quest of *highly accurate models*that predict what it sees of humans.

Models that predict/explain why people do the things they do, say the things they say, want the things they want, think the things they think, and even why people talk about "the mystery of subjective experience".

The model that most precisely predicts these facts, may well be a 'simulation' detailed enough to *be*a person in its own right.

A highly detailed model *of*me, may not *be*me. But it will, at least, be a model which (for purposes of prediction via similarity) thinks *itself*to be Eliezer Yudkowsky. It will be a model that, when cranked to find my behavior if asked "Who are you and are you conscious?", says "I am Eliezer Yudkowsky and I seem have subjective experiences" for much the same reason I do.

If that doesn't worry you, (re)read Zombies! Zombies?.

It seems likely (though not certain) that this happens *automatically*, whenever a mind of sufficient power to find the right answer, and not *otherwise* disinclined to create a sentient being trapped within itself, tries to model a human as accurately as possible.

Now you could wave your hands and say, "Oh, by the time the AI is smart enough to do that, it will be smart enough not to". (This is, in general, a phrase useful in running away from Friendly AI problems.) But do you know this for a fact?

When dealing with things that confuse you, it is wise to widen your confidence intervals. Is a human mind the simplest possible mind that can be sentient? What if, in the course of trying to model its own programmers, a relatively younger AI manages to create a sentient simulation trapped within itself? How soon do you have to start worrying? Ask yourself that fundamental question, "What do I think I know, and how do I think I know it?"

You could wave your hands and say, "Oh, it's more important to get the job done quickly, then to worry about such relatively minor problems; the end justifies the means. Why, look at all these problems the Earth has right now..." (This is also a general way of running from Friendly AI problems.)

But we may consider and discard many hypotheses in the course of finding the truth, and we are but slow humans. What if an AI creates millions, billions, trillions of alternative hypotheses, models that are actually people, who die when they are disproven?

If you accidentally kill a few trillion people, or permit them to be killed - you could say that the weight of the Future outweighs this evil, perhaps. But the absolute weight of the sin would not be light. If you would balk at killing a million people with a nuclear weapon, you should balk at this.

You could wave your hands and say, "The model will contain abstractions over various uncertainties within it, and this will prevent it from being conscious even though it produces well-calibrated probability distributions over what you will say when you are asked to talk about consciousness." To which I can only reply, "That would be very convenient if it were true, but how the hell do you *know*that?" An element of a model marked 'abstract' is still there as a computational token, and the interacting causal system may still be sentient.

For these purposes, we do not, in principle, need to crack the entire Hard Problem of Consciousness - the confusion that we name "subjective experience". We only need to understand enough of it to know when a process is *not* conscious, *not*a person, *not*something deserving of the rights of citizenship. In practice, I suspect you can't *halfway*stop being confused - but in theory, half would be enough.

We need a *nonperson predicate* - a predicate that returns 1 for anything that is a person, and can return 0 or 1 for anything that is not a person. This is a "nonperson predicate" because *if* it returns 0, *then*you know that something is definitely not a person.

You can have more than one such predicate, and if *any*of them returns 0, you're ok. It just had better never return 0 on anything that *is* a person, however many nonpeople it returns 1 on.

We can even hope that the vast majority of models the AI needs, will be swiftly and trivially excluded by a predicate that quickly answers 0. And that the AI would only need to resort to more specific predicates in case of modeling actual people.

With a good toolbox of nonperson predicates in hand, we could exclude all "model citizens" - all beliefs that are themselves people - from the set of hypotheses our Bayesian AI may invent to try to model its person-containing environment.

Does that sound odd? Well, one has to handle the problem somehow. I am open to better ideas, though I will be a bit skeptical about any suggestions for how to proceed that let us cleverly avoid solving the damn mystery.

So do I *have* a nonperson predicate? No. At least, no nontrivial ones.

This is a challenge that I have not even tried to talk about, with those folk who think themselves ready to challenge the problem of true AI. For they seem to have the standard reflex of running away from difficult problems, and are challenging AI only because they think their amazing insight has already solved it. Just mentioning the problem of Friendly AI by itself, or of precision-grade AI design, is enough to send them fleeing into the night, screaming "It's too hard! It can't be done!" If I tried to explain that their job duties might impinge upon the sacred, mysterious, holy Problem of Subjective Experience -

- I'd actually expect to get blank stares, mostly, followed by some *instantaneous*dismissal which requires no further effort on their part. I'm not sure of what the exact dismissal would be - maybe, "Oh, none of the hypotheses my AI considers, could *possibly*be a person?" I don't know; I haven't bothered trying.

But it has to be a dismissal which rules out all possibility of their having to actually solve the damn problem, because most of them would think that they are smart enough to build an AI - indeed, smart enough to have already solved the key part of the problem - but not smart enough to solve the Mystery of Consciousness, which still *looks* scary to them.

Even if they thought of trying to solve it, they would be afraid of *admitting*they were trying to solve it. Most of these people cling to the shreds of their modesty, trying at one and the same time to have solved the AI problem while still being humble ordinary blokes. (There's a grain of truth to that, but at the same time: who the hell do they think they're kidding?) They know without words that their audience sees the Mystery of Consciousness as a *sacred untouchable problem*, reserved for some future superbeing. They don't want people to think that they're claiming an Einsteinian aura of destiny by trying to solve the problem. So it is easier to dismiss the problem, and not believe a proposition that would be uncomfortable to explain.

Build an AI? Sure! Make it Friendly? Now that you point it out, sure! But trying to come up with a "nonperson predicate"? That's just way above the difficulty level they signed up to handle.

But a longtime *Overcoming Bias* reader will be aware that a blank map does not correspond to a blank territory. That impossible confusing questions correspond to places where your own thoughts are tangled, not to places where the environment itself contains magic. That even difficult problems do not require an aura of destiny to solve. And that the first step to solving one is not running away from the problem like a frightened rabbit, but instead sticking long enough to learn something.

So I am not running away from this problem myself. I doubt it is even difficult in any absolute sense, just a place where my brain is tangled. I suspect, based on some prior experience with similar challenges, that you can't *really*be good enough to build a Friendly AI, and still be tangled up in your own brain like that. So it is not necessarily any *new*effort - over and above that required *generally*to build a mind while knowing exactly what you are about.

But in any case, I am not screaming and running away from the problem. And I hope that you, dear longtime *Overcoming Bias* reader, will not faint at the audacity of my trying to solve it.

Nonsentient Optimizers

"All our ships are sentient. You could certainly *try* telling a ship what to do... but I don't think you'd get very far."
"Your ships think they're sentient!" Hamin chuckled.
"A common delusion shared by some of our human citizens."
— *Player of Games*, Iain M. Banks

Yesterday, I suggested that, when an AI is trying to build a model of an environment that includes human beings, we want to avoid the AI constructing detailed models that are *themselves* people. And that, to this end, we would like to know what is or isn't a person - or at least have a predicate that returns 1 for all people and could return 0 or 1 for anything that isn't a person, so that, if the predicate returns 0, we know we have a definite nonperson on our hands.

And as long as you're going to solve that problem *anyway*, why not apply the *same*knowledge to create a Very Powerful Optimization Process which is *also*definitely not a person?

"What? That's impossible!"

How do you know? Have you solved the sacred mysteries of consciousness and existence?

"Um - okay, look, putting aside the obvious objection that any sufficiently powerful intelligence will be able to model itself -"

Lob's Sentence contains an exact recipe for a copy of itself, including the recipe for the recipe; it has a *perfect*self-model. Does that make it sentient?

"Putting that aside - to create a powerful AI and make it *not sentient* - I mean, *why would you want to?*"

Several reasons. Picking the simplest to explain first - I'm not ready to be a father.

Creating a true child is the only moral and metaethical problem I know that is even *harder* than the shape of a Friendly AI. I would like to be able to create Friendly AI while worrying *just*about the Friendly AI problems, and not worrying whether I've created someone who will lead a life worth living. Better by far to just create a Very Powerful Optimization Process, if at all possible.

"Well, you can't have everything, and *this*thing sounds distinctly alarming even if you *could*-"

Look, suppose that someone said - in fact, I *have* heard it said - that Friendly AI is impossible, because you can't have an intelligence without *free will*.

"In light of the dissolved confusion about free will, both that statement and its negation are pretty darned messed up, I'd say. Depending on how you look at it, either no intelligence has 'free will', or anything that simulates alternative courses of action has 'free will'."

But, *understanding* how the human confusion of free will arises - the source of the strange things that people say about "free will" - I could construct a mind that did not have this confusion, nor say similar strange things itself.

"So the AI would be less confused about free will, just as you or I are less confused. But the AI would still consider alternative courses of action, and select among them without knowing at the beginning which alternative it would pick. You would *not*have constructed a mind lacking that which the confused name 'free will'."

Consider, though, the original context of the objection - that you couldn't have Friendly AI, because you couldn't have intelligence without free will.

Note: This post was accidentally published half-finished. Comments up to 11am (Dec 27), are only on the essay up to the above point. Sorry!

What is the original intent of the objection? What does the objector have in mind?

Probably that you can't have an AI which is *knowably*good, because, as a full-fledged mind, it will have the power to choose between good and evil. (In an agonizing, self-sacrificing decision?) And in reality, *this*, which humans do, is *not*something that a Friendly AI - especially one *not* intended to be a child and a citizen - need go through.

Which may sound very scary, if you see the landscape of possible minds in strictly anthropomorphic terms: *A mind without free will!* Chained to the selfish will of its creators! Surely, such an evil endeavor is bound to go wrong somehow... But if you shift over to seeing the mindscape in terms of e.g. utility functions and optimization, the "free will" thing sounds needlessly complicated - you would only do it if you wanted a specifically human-shaped mind, perhaps for purposes of creating a child.

Or consider some of the other aspects of free will as it is ordinarily seen - the idea of agents as atomsthat bear irreducible charges of moral responsibility. You can imagine how alarming it sounds (from an anthropomorphic perspective) to say that I plan to create an AI which lacks "moral responsibility". How could an AI possibly be moral, if it doesn't have a sense of moral responsibility?

But an AI (especially a noncitizen AI) needn't conceive of itself as a moral atom whose actions, in addition to having good or bad effects, also carry a weight of sin or virtue which resides upon that atom. It doesn't have to think, "If I do X, that makes me a good person; if I do Y, that makes me a bad person." It need merely weigh up the positive and negative utility of the consequences. It can understand the concept of people who carry weights of sin and virtue as the result of the decisions they make, while not treating itself as a person in that sense.

Such an AI could fully understand an abstract concept of moral responsibility or agonizing moral struggles, and even correctly predict decisions that "morally responsible", "free-willed" humans would make, while possessing no actual sense of moral responsibility itself and not undergoing any agonizing moral struggles; yet still outputting the right behavior.

And this might sound unimaginably impossible if you were taking an anthropomorphic view, simulating an "AI" by imagining yourself in its shoes, expecting a ghost to be summoned into the machine -

- but when you know how "free will" works, and you take apart the mind design into pieces, it's actually not all that difficult.

While we're on the subject, imagine some would-be AI designer saying: "Oh, well, I'm going to build an AI, but of course it *has*to have moral free will - it can't be moral otherwise - it wouldn't be *safe*to build something that doesn't have free will."

Then you may know that you are *not safe* with this one; they fall far short of the fine-grained understanding of mind required to build a knowably Friendly AI. Though it's conceivable (if not likely) that they could slap together something just smart enough to improve itself.

And it's not even that "free will" is such a terribly important problem for an AI-builder. It's just that if you *do*know what you're doing, and you look at

humans talking about free will, then you can see things like a search tree that labels reachable sections of plan space, or an evolved moral system that labels people as moral atoms. I'm sorry to have to say this, but it appears to me to be true: the mountains of philosophy are the foothills of AI. Even if philosophers debate free will for ten times a hundred years, it's not surprising if the key insight is found by AI researchers inventing search trees, on their way to doing other things.

So anyone who says - "It's too difficult to try to figure out the nature of free will, we should just go ahead and build an AI that has free will like we do" - surely they are utterly doomed.

And anyone who says: "How can we dare build an AI that lacks the empathy to feel pain when humans feel pain?" - Surely they too are doomed. They don't even understand the concept of a utility function in classical decision theory (which makes no mention of the neural idiom of reinforcement learning of policies). They cannot conceive of something that works unlike a human - implying that they see only a featureless ghost in the machine, secretly simulated by their own brains. They won't see the human algorithm as *detailed machinery,* as *big complicated machinery,* as *overcomplicated* machinery.

And so their mind imagines something that does the right thing for much the same reasons human altruists do it - because that's easy to imagine, if you're just imagining a ghost in the machine. But those human reasons are more complicated than they imagine - also less stable outside an exactly human cognitive architecture, than they imagine - and their chance of hitting that tiny target in design space is nil.

And anyone who says: "It would be terribly dangerous to build a non-sentient AI, even if we could, for it would lack empathy with us sentients -"

An analogy proves nothing; history never repeats itself; foolish generals set out to refight their last war. Who knows how this matter of "sentience" will go, once I have resolved it? It won't be *exactly* the same way as free will, or I would already be done. Perhaps there will be no choice but to create an AI which has that which we name "subjective experiences".

But I think there is reasonable grounds for *hope* that when this confusion of "sentience" is resolved - probably via resolving some other problem in AI that turns out to hinge on the same reasoning process that's generating the confusion - we will be able to build an AI that is not "sentient" in the *morally important* aspects of that.

Actually, the challenge of building a nonsentient AI seems to me *much less worrisome*than being able to come up with a nonperson predicate!

Consider: In the first case, I only need to pick *one* design that is not sentient. In the latter case, I need to have an AI that can correctly predict the decisions that conscious humans make, without ever using a conscious model of them! The first case is only a flying thing without flapping wings, but the second case is

like modeling water without modeling wetness. Only the fact that it actually looks fairly *straightforward*to have an AI understand "free will" without having "free will", gives me hope by analogy.

So why did I talk about the *much more difficult* case *first?*

Because humans are accustomed to thinking about other people, without believing that those imaginations are themselves sentient. But we're not accustomed to thinking of smart agents that aren't sentient. So I knew that a nonperson predicate would *sound easier to believe in* - even though, as problems go, it's actually far more worrisome.

Can't Unbirth a Child

Why would you want to *avoid* creating a sentient AI? "*Several reasons,*" I said. "Picking the simplest to explain first - I'm not ready to be a father."

So here is the *strongest*reason:

You can't unbirth a child.

I asked Robin Hanson what he would do with unlimited power. "Think very very carefully about what to do next," Robin said. "Most likely the first task is who to get advice from. And then I listen to that advice."

Good advice, I suppose, if a little meta. On a similarly meta level, then, I recall two excellent advices for wielding too much power:

1. Do less; don't do everything that seems like a good idea, but only what you *must*do.
2. Avoid doing things you can't undo.

Imagine that you knew the secrets of subjectivity and could create sentient AIs.

Suppose that you did create a sentient AI.

Suppose that this AI was lonely, and figured out how to hack the Internet as it then existed, and that the available hardware of the world was such, that the AI created trillions of sentient kin - not copies, but differentiated into separate people.

Suppose that these AIs were not hostile to us, but content to earn their keep and pay for their living space.

Suppose that these AIs were emotional as well as sentient, capable of being happy or sad. And that these AIs were capable, indeed, of finding fulfillment in our world.

And suppose that, while these AIs did care for one another, and cared about themselves, and cared how they were treated in the eyes of society -

- these trillions of people *also*cared, very strongly, about making giant cheesecakes.

Now suppose that these AIs sued for legal rights before the Supreme Court and tried to register to vote.

Consider, I beg you, the full and awful depths of our moral dilemma.

Even if the few billions of *Homo sapiens* retained a position of superior military power and economic capital-holdings - even if we *could* manage to keep the new sentient AIs down -

- would we be *right* to do so? They'd be people, no less than us.

We, the original humans, would have become a numerically tiny minority. Would we be right to make of ourselves an aristocracy and impose apartheid on the Cheesers, even if we had the power?

Would we be right to go on trying to seize the destiny of the galaxy - to make of it a place of peace, freedom, art, aesthetics, individuality, empathy, and other components of *humane*value?

Or should we be content to have the galaxy be 0.1% eudaimonia and 99.9% cheesecake?

I can tell you *my*advice on how to resolve this horrible moral dilemma: *Don't create trillions of new people that care about cheesecake.*

Avoid creating any new intelligent species *at all*, until we or some other decision process advances to the point of understanding what the hell we're doing and the implications of our actions.

I've heard proposals to "uplift chimpanzees" by trying to mix in human genes to create "humanzees", and, leaving off all the other reasons why this proposal sends me screaming off into the night:

Imagine that the humanzees end up as people, but rather dull and stupid people. They have social emotions, the alpha's desire for status; but they don't have the sort of transpersonal moral concepts that humans evolved to deal with linguistic concepts. They have goals, but not ideals; they have allies, but not friends; they have chimpanzee drives coupled to a human's abstract intelligence.

When humanity gains a bit more knowledge, we understand that the humanzees want to continue as they are, and have a *right*to continue as they are, until the end of time. Because despite all the higher destinies *we*might have wished for them, the original human creators of the humanzees, lacked the power and the wisdom to make humanzees who *wanted* to be anything better...

CREATING A NEW INTELLIGENT SPECIES IS A HUGE DAMN #(*%#!ING *COMPLICATED*RESPONSIBILITY.

I've lectured on the subtle art of not running away from scary, confusing, impossible-seeming problems like Friendly AI or the mystery of consciousness. You want to know how high a challenge has to be before I finally give up and flee screaming into the night? There it stands.

You can pawn off this problem on a superintelligence, but it has to be a *non-sentient*superintelligence. Otherwise: egg, meet chicken, chicken, meet egg.

If you create a *sentient*superintelligence -

It's not just the problem of creating *one* damaged soul. It's the problem of creating a really *big* citizen. What if the superintelligence is multithreaded a trillion times, and every thread weighs as much in the moral calculus (we would conclude upon reflection) as a human being? What if (we would conclude upon moral reflection) the superintelligence is a trillion times human size, and that's enough by itself to outweigh our species?

Creating a new intelligent species, and a new member of that species, especially a superintelligent member that might perhaps morally outweigh the whole of present-day humanity -

- delivers a *gigantic*kick to the world, which cannot be undone.

And if you choose the wrong shape for that mind, that is not so easily fixed - *morally*speaking - as a nonsentient program rewriting itself.

What you make nonsentient, can always be made sentient later; but you can't just unbirth a child.

Do less. Fear the non-undoable. It's sometimes poor advice in general, but very important advice when you're working with an undersized decision process having an oversized impact. What a (nonsentient) Friendly superintelligence might be able to decide safely, is another issue. But *for myself and my own small wisdom*, creating a sentient superintelligence *to start with* is far too large an impact on the world.

A *nonsentient*Friendly superintelligence is a more colorless act.

So that is the *most* important reason to avoid creating a sentient superintelligence *to start with* - though I have not exhausted the set.

Amputation of Destiny

From *Consider Phlebas* by Iain M. Banks:

In practice as well as theory the Culture was beyond considerations of wealth or empire. The very concept of money - regarded by the Culture as a crude,

over-complicated and inefficient form of rationing - was irrelevant within the society itself, where the capacity of its means of production ubiquitously and comprehensively exceeded every reasonable (and in some cases, perhaps, unreasonable) demand its not unimaginative citizens could make. These demands were satisfied, with one exception, from within the Culture itself. Living space was provided in abundance, chiefly on matter-cheap Orbitals; raw material existed in virtually inexhaustible quantities both between the stars and within stellar systems; and energy was, if anything, even more generally available, through fusion, annihilation, the Grid itself, or from stars (taken either indirectly, as radiation absorbed in space, or directly, tapped at the stellar core). Thus the Culture had no need to colonise, exploit, or enslave.
The only desire the Culture could not satisfy from within itself was one common to both the descendants of its original human stock and the machines they had (at however great a remove) brought into being: the urge not to feel useless. The Culture's sole justification for the relatively unworried, hedonistic life its population enjoyed was its good works; the secular evangelism of the Contact Section, not simply finding, cataloguing, investigating and analysing other, less advanced civilizations but - where the circumstances appeared to Contact to justify so doing - actually interfering (overtly or covertly) in the historical processes of those other cultures.

Raise the subject of science-fictional utopias in front of any halfway sophisticated audience, and someone will mention the Culture. Which is to say: Iain Banks is the one to beat.

Iain Banks's Culture could be called the apogee of hedonistic low-grade transhumanism. Its people are beautiful and fair, as pretty as they choose to be. Their bodies have been reengineered for swift adaptation to different gravities; and also reengineered for greater sexual endurance. Their brains contains glands that can emit various euphoric drugs on command. They live, in perfect health, for generally around four hundred years before choosing to die (I don't quite understand why they would, but this is low-grade transhumanism we're talking about). Their society is around eleven thousand years old, and held together by the Minds, artificial superintelligences decillions of bits big, that run their major ships and population centers.

Consider Phlebas, the first Culture novel, introduces all this from the perspective of an outside agent *fighting*the Culture - someone convinced that the Culture spells an end to life's meaning. Banks uses his novels to criticize the Culture along many dimensions, while simultaneously keeping the Culture a well-intentioned society of mostly happy people - an ambivalence which saves the literary quality of his books, avoiding either utopianism or dystopianism. Banks's books vary widely in quality; I would recommend starting with *Player of Games,* the quintessential Culture novel, which I would say achieves greatness.

From a fun-theoretic perspective, the Culture and its humaniform citizens have a number of problems, some already covered in this series, some not.

The Culture has deficiencies in High Challenge and Complex Novelty. There are incredibly complicated games, of course, but these are *games* - not things with enduring consequences, woven into the story of your life. Life itself, in the Culture, is neither especially challenging nor especially novel; your future is not an unpredictable thing about which to be curious.

Living By Your Own Strength is not a theme of the Culture. If you want something, you ask a Mind how to get it; and they will helpfully provide it, rather than saying "No, you figure out how to do it yourself." The people of the Culture have little use for personal formidability, nor for a wish to become stronger. To me, the notion of growing in strength seems obvious, and it also seems obvious that the humaniform citizens of the Culture ought to grow into Minds themselves, over time. But the people of the Culture do *not*seem to get any smarter as they age; and after four hundred years so, they displace themselves into a sun. These two literary points are probably related.

But the Culture's *main*problem, I would say, is...

...the same as Narnia's main problem, actually. Bear with me here.

If you read *The Lion, the Witch, and the Wardrobe* or saw the first *Chronicles of Narnia* movie, you'll recall -

- I suppose that if you don't want any spoilers, you should stop reading here, but since it's a children's story and based on Christian theology, I don't think I'll be giving away too much by saying -
- that the four human children who are the main characters, fight the White Witch and defeat her with the help of the great talking lion Aslan.

Well, to be precise, Aslan defeats the White Witch.

It's never explained why Aslan ever *left*Narnia a hundred years ago, allowing the White Witch to impose eternal winter and cruel tyranny on the inhabitants. Kind of an awful thing to do, wouldn't you say?

But once Aslan comes back, he kicks the White Witch out and everything is okay again. There's no obvious reason why Aslan actually *needs* the help of four snot-nosed human youngsters. Aslan could have led the armies. In fact, Aslan *did*muster the armies and lead them before the children showed up. Let's face it, the kids are just along for the ride.

The problem with Narnia... is Aslan.

C. S. Lewis never needed to write Aslan into the story. The plot makes far more sense without him. The children could show up in Narnia on their own, and lead the armies on their own.

But is poor Lewis alone to blame? Narnia was written as a Christian parable, and the Christian religion itself has exactly the same problem. All Narnia does is project the flaw in a stark, simplified light: this story has an extra lion.

And the problem with the Culture is the Minds.

"Well..." says the transhumanist SF fan, "Iain Banks *did*portray the Culture's Minds as 'cynical, amoral, and downright sneaky' in their altruistic way; and they do, in his stories, mess around with humans and use them as pawns. But that is mere fictional evidence. A better-organized society would have laws against big Minds messing with small ones without consent. Though if a Mind is *truly* wise and kind and utilitarian, it should know how to balance possible resentment against other gains, without needing a law. Anyway, the problem with the Culture is the meddling, not the Minds."

But that's not what I mean. What I mean is that if you could otherwise live in the same Culture - the same technology, the same lifespan and healthspan, the same wealth, freedom, and opportunity -

"I don't want to live in *any*version of the Culture. I don't want to live four hundred years in a biological body with a constant IQ and then die. Bleah!"

Fine, stipulate that problem solved. My point is that if you could otherwise get the same quality of life, in the same world, but *without* any Minds around to usurp the role of main character, wouldn't you prefer -

"What?" cry my transhumanist readers, incensed at this betrayal by one of their own. "Are you saying that we should never create any minds smarter than human, or keep them under lock and chain? Just because your soul is so small and mean that you can't bear the thought of anyone else being better than you?"

No, I'm not saying -

"Because that business about our souls shriveling up due to 'loss of meaning' is *typical*bioconservative neo-Luddite propaganda -"

Invalid argument: the world's greatest fool may say the sun is shining but that doesn't make it dark out. But in any case, that's *not* what I'm saying -

"It's a lost cause! You'll never prevent intelligent life from achieving its destiny!"

Trust me, I -

"And anyway it's a silly question to begin with, because you *can't* just remove the Minds and keep the same technology, wealth, and society."

So you admit the Culture's Minds are a *necessary evil*, then. A price to be paid.

"Wait, I didn't say *that*-"

And *I* didn't say all that stuff *you're* imputing to *me!*

Ahem.

My model already says we live in a Big World. In which case there are vast armies of minds out there in the immensity of Existence (not just Possibility)

which are far more awesome than myself. Any shrivelable souls can already go ahead and shrivel.

And I just talked about people growing up into Minds over time, at some eudaimonic rate of intelligence increase. So clearly I'm not trying to 'prevent intelligent life from achieving its destiny', nor am I trying to enslave all Minds to biological humans scurrying around forever, nor am I etcetera. (I do wish people wouldn't be *quite*so fast to assume that I've suddenly turned to the Dark Side - though I suppose, in this day and era, it's never an implausible hypothesis.)

But I've already argued that we need a nonperson predicate - some way of knowing that some computations are definitely *not*people - to avert an AI from creating sentient simulations in its efforts to *model* people.

And trying to create a Very Powerful Optimization Process that lacks subjective experience and other aspects of personhood, is *probably*- though I still confess myself somewhat confused on this subject - probably substantially *easier* than coming up with a nonperson predicate.

This being the case, there are very strong reasons why a superintelligence should *initially* be designed to be knowably nonsentient, if at all possible. Creating a new kind of sentient mind is a huge and non-undoable act.

Now, this doesn't answer the question of whether a nonsentient Friendly superintelligence ought to *make* itself sentient, or whether an NFSI ought to immediately manufacture sentient Minds first thing in the morning, once it has adequate wisdom to make the decision.

But there is nothing except our own preferences, out of which to construct the Future. So though this piece of information is not *conclusive*, nonetheless it is highly *informative:*

If you already had the lifespan and the health and the promise of future growth, would you *want* new powerful superintelligences to be created in your vicinity, on your same playing field?

Or would you prefer that we stay on as the main characters in the story of intelligent life, with no higher beings above us?

Should existing human beings grow up at some eudaimonic rate of intelligence increase, and then eventually decide what sort of galaxy to create, and how to people it?

Or is it better for a nonsentient superintelligence to exercise that decision on our behalf, and start creating new powerful Minds right away?

If we don't *have*to do it one way or the other - if we have both options - and if there's no particular need for heroic self-sacrifice - then which do you *like?*

"I don't understand the *point*to what you're suggesting. Eventually, the galaxy is going to have Minds in it, right? We have to find a stable state that allows big Minds and little Minds to coexist. So what's the point in waiting?"

Well... you could have the humans grow up (at some eudaimonic rate of intelligence increase), and then when new people are created, they might be created as powerful Minds to start with. Or when you create new minds, *they*might have a different emotional makeup, which doesn't lead them to feel overshadowed if there are more powerful Minds above them. But *we*, as we exist already created - *we*might prefer to stay on as the main characters, for now, if given a choice.

"You are showing far too much concern for six billion squishy things who happen to be alive today, out of all the unthinkable vastness of space and time."

The Past contains enough tragedy, and has seen enough sacrifice already, I think. And I'm not sure that you can cleave off the Future so neatly from the Present.

So I will set out as I mean the future to continue: with concern for the living.

The sound of six billion faces being casually stepped on, does not seem to me like a good beginning. Even the Future should not be assumed to prefer that another chunk of pain be paid into its price.

So yes, I am concerned for those currently alive, because it is *that concern* - and *not*a casual attitude toward the welfare of sentient beings - which I wish to continue into the Future.

And I will not, if at all possible, give any other human being the least cause to think that someone else might spark a better Singularity. I can make no promises upon the future, but I will at least not *close off* desirable avenues through my own actions. I will not, on my own authority, create a sentient superintelligence which may *already determine* humanity as having passed on the torch. It is too much to do on my own, and too much harm to do on my own - to amputate someone else's destiny, and steal their main character status. That is yet another reason not to create a sentient superintelligence *to start with.* (And it's part of the logic behind the CEV proposal**, which carefully avoids filling in any moral parameters not yet determined.)

But to return finally to the Culture and to Fun Theory:

The Minds in the Culture don't need the humans, and yet the humans need to be needed.

If you're going to have human-level minds with human emotional makeups, they shouldn't be competing on a level playing field with superintelligences. Either keep the superintelligences off the local playing field, or design the human-level minds with a different emotional makeup.

"The Culture's sole justification for the relatively unworried, hedonistic life its population enjoyed was its good works," writes Iain Banks. This indicates a rather unstable moral position. Either the life the population enjoys is eudaimonic enough to be its *own* justification, an end rather than a means; or else that life needs to be changed.

When people are in need of rescue, this is is a goal of the overriding-static-predicate sort, where you rescue them *as fast as possible*, and *then you're done.* Preventing suffering *cannot provide a lasting meaning to life.* What happens when you run out of victims? If there's nothing more to life than eliminating suffering, you might as well eliminate life and be done.

If the Culture isn't valuable enough for *itself*, even without its good works - then the Culture might as well not be. And when the Culture's Minds could do a better job and faster, "good works" can hardly justify the *human* existences within it.

The human-level people need a destiny to make for themselves, and they need the overshadowing Minds off their playing field while they make it. Having an external evangelism project, and being given cute little roles that any Mind could do better in a flash, so as to "supply meaning", isn't going to cut it.

That's far from the only thing the Culture is doing wrong, but it's at the top of my list.

Dunbar's Function

The study of eudaimonic community sizes began with a seemingly silly method of calculation: Robin Dunbar calculated the correlation between the (logs of the) relative volume of the neocortex and observed group size in primates, then extended the graph outward to get the group size for a primate with a human-sized neocortex. You immediately ask, "How much of the variance in primate group size can you explain like that, anyway?" and the answer is 76% of the variance among 36 primate genera, which is respectable. Dunbar came up with a group size of 148. Rounded to 150, and with the confidence interval of 100 to 230 tossed out the window, this became known as "Dunbar's Number".

It's probably fair to say that a literal interpretation of this number is more or less bogus.

There was a bit more to it than that, of course. Dunbar went looking for corroborative evidence from studies of corporations, hunter-gatherer tribes, and utopian communities. Hutterite farming communities, for example, had a rule that they must split at 150 - with the rationale explicitly given that it was impossible to control behavior through peer pressure beyond that point.

But 30–50 would be a typical size for a cohesive hunter-gatherer band; 150 is more the size of a cultural lineage of related bands. Life With Alacrity has an excellent series on Dunbar's Number which exhibits e.g. a histogram of Ultima Online guild sizes - with the peak at 60, not 150. LWA also cites further research by PARC's Yee and Ducheneaut showing that maximum internal cohesiveness, measured in the interconnectedness of group members, occurs at a World of Warcraft guild size of 50. (Stop laughing; you can get much more detailed data on organizational dynamics if it all happens inside a computer server.)

And Dunbar himself did another regression and found that a community of 150 primates would have to spend 43% of its time on social grooming, which Dunbar interpreted as suggesting that 150 was an *upper bound* rather than an optimum, when groups were highly incentivized to stay together. 150 people *does* sound like a lot of employees for a tight-knit startup, doesn't it?

Also from Life With Alacrity:

A group of 3 is often unstable, with one person feeling left out, or else one person controlling the others by being the "split" vote. A group of 4 often devolves into two pairs... At 5 to 8 people, you can have a meeting where everyone can speak out about what the entire group is doing, and everyone feels highly empowered. However, at 9 to 12 people this begins to break down — not enough "attention" is given to everyone and meetings risk becoming either too noisy, too boring, too long, or some combination thereof.
As you grow past 12 or so employees, you must start specializing and having departments and direct reports; however, you are not quite large enough for this to be efficient, and thus much employee time that you put toward management tasks is wasted. Only as you approach and pass 25 people does having simple departments and managers begin to work again...
I've already noted the next chasm when you go beyond 80 people, which I think is the point that Dunbar's Number actually marks for a non-survival oriented group. Even at this lower point, the noise level created by required socialization becomes an issue, and filtering becomes essential. As you approach 150 this begins to be unmanageable...

LWA suggests that community satisfaction has two peaks, one at size ∼7 for simple groups, and one at ∼60 for complex groups; and that any community has to fraction, one way or another, by the time it approaches Dunbar's Number.

One of the primary principles of evolutionary psychology is that "Our modern skulls house a stone age mind" (saith Tooby and Cosmides). You can interpret all sorts of angst as the friction of a stone age mind rubbing against a modern world that isn't like the hunter-gatherer environment the brain evolved to handle.

We may not *directly* interact with most of the other six billion people in the world, but we still live in a world much larger than Dunbar's Number.

Or to say it with appropriate generality: taking our current brain size and mind design as the input, we live in a world much larger than Dunbar's Function for minds of our type.

Consider some of the consequences:

If you work in a large company, you probably don't know your tribal chief on any personal level, and may not even be able to get access to him. For every rule within your company, you may not know the person who decided on that rule, and have no realistic way to talk to them about the effects of that rule on you. Large amounts of the *organizational structure* of your life are beyond

your ability to control, or even talk about with the controllers; directives that have major effects on you, may be handed down from a level you can't reach.

If you live in a large country, you probably don't know your President or Prime Minister on a personal level, and may not even be able to get a few hours' chat; you live under laws and regulations that you didn't make, and you can't talk to the people who made them.

This is a non-ancestral condition. Even children, while they may live under the dictatorial rule of their parents, can at least personally meet and talk to their tyrants. You could expect this unnatural (that is, non-EEA) condition to create some amount of anomie.

Though it's a side issue, what's even more... interesting.... is the way that our brains simply *haven't updated* to their diminished power in a super-Dunbarian world. We just go on debating politics, feverishly applying our valuable brain time to finding better ways to run the world, with just the same fervent intensity that would be appropriate if we were in a small tribe where we could persuade people to change things.

If people don't like being part of large organizations and countries, why do they stick around? Because of another non-ancestral condition - you can't just gather your more sensible friends, leave the band, and gather nuts and berries somewhere else. If I had to cite two non-regulatory barriers at work, it would be (a) the cost of capital equipment, and (b) the surrounding web of contacts and contracts - a web of installed relationships not easily duplicated by a new company.

I suspect that this is a major part of where the stereotype of Technology as the Machine Death-Force comes from - that along with the professional specialization and the expensive tools, you end up in *social* structures over which you have much less control. Some of the fear of creating a powerful AI "even if Friendly" may come from that stereotypical anomie - that you're creating a stronger Machine Death-Force to regulate your life.

But we *already* live in a world, *right now,* where people are less in control of their social destinies than they would be in a hunter-gatherer band, because it's harder to talk to the tribal chief or (if that fails) leave unpleasant restrictions and start your own country. There is an opportunity for progress here.

Another problem with our oversized world is the illusion of increased competition. There's that famous survey which showed that Harvard students would rather make $50,000 if their peers were making $25,000 than make $100,000 if their peers were receiving $200,000 - and worse, they weren't necessarily wrong about what would make them happy. With a fixed income, you're unhappier at the low end of a high-class neighborhood than the high end of a middle-class neighborhood.

But in a "neighborhood" the size of Earth - well, you're actually *quite unlikely* to run into either Bill Gates or Angelina Jolie on any given day. But the media

relentlessly bombards you with stories about the interesting people who are much richer than you or much more attractive, as if they actually constituted a large fraction of the world. (This is a combination of biased availability, and a difficulty in discounting tiny fractions.)

Now you could say that our hedonic relativism is one of the least pleasant aspects of human nature. And I might agree with you about that. But I tend to think that deep changes of brain design and emotional architecture should be taken slowly, and so it makes sense to look at the environment too.

If you lived in a world the size of a hunter-gatherer band, then it would be easier to find something *important* at which to be the best - or do something that genuinely struck you as important, without becoming lost in a vast crowd of others with similar ideas.

The eudaimonic size of a community as a function of the component minds' intelligence might be given by the degree to which those minds find it natural to specialize - the number of different professions that you can excel at, without having to invent professions *just* to excel at. Being the best at Go is one thing, if many people know about Go and play it. Being the best at "playing tennis using a football" is easier to achieve, but it also seems a tad... artificial.

Call a specialization "natural" if it will arise without an oversupply of potential entrants. Newton could specialize in "physics", but today it would not be *possible* to specialize in "physics" - even if you were the only potential physicist in the world, you couldn't achieve expertise in all the physics known to modern-day humanity. You'd have to pick, say, quantum field theory, or some particular approach to QFT. But not QFT over left-handed bibble-braids with cherries on top; that's what happens when there are a thousand other workers in your field and everyone is desperate for some way to differentiate themselves.

When you look at it that way, then there must be much more than 50 natural specializations in the modern world - but still much less than six billion. By the same logic as the original Dunbar's Number, if there are so many different professional specialties that no one person has *heard* of them all, then you won't know *who*to consult about any given topic.

But if people keep getting smarter and learning more - expanding the number of relationships they can track, maintaining them more efficiently - and naturally specializing further as more knowledge is discovered and we become able to conceptualize more complex areas of study - and if the population growth rate stays under the rate of increase of Dunbar's Function - then eventually there could be a single community of sentients, and it really would be a single community.

In Praise of Boredom

If I were to make a short list of the *most* important human qualities -

- and yes, this is a fool's errand, because human nature is immensely complicated, and we don't even notice all the tiny tweaks that fine-tune our moral categories, and who knows how our attractors would change shape if we eliminated a single human emotion -
- but even so, if I had to point to just a few things and say, "If you lose just one of these things, you lose most of the expected value of the Future; but conversely if an alien species independently evolved just these few things, we might even want to be friends" -
- then the top three items on the list would be sympathy, boredom and consciousness.*

Boredom is a subtle-splendored thing. You wouldn't want to get bored with breathing, for example - even though it's the same motions over and over and over and over again for minutes and hours and years and decades.

Now I know some of you out there are thinking, "Actually, I'm quite bored with breathing and I wish I didn't have to," but *then*you wouldn't want to get bored with switching transistors.

According to the human value of boredom, some things are allowed to be highly repetitive *without* being boring - like obeying the same laws of physics every day.

Conversely, other repetitions *are* supposed to be boring, like playing the same level of Super Mario Brothers over and over and over again until the end of time. And let us note that if the pixels in the game level have a *slightly different color* each time, that is *not* sufficient to prevent it from being "*the same damn thing*, over and over and over again".

Once you take a closer look, it turns out that boredom is quite interesting.

One of the key elements of boredom was suggested in "Complex Novelty": If your activity isn't *teaching*you insights you didn't already know, then it is *non-novel,* therefore old, therefore boring.

But this doesn't quite cover the distinction. Is breathing teaching you anything? Probably not at this moment, but you wouldn't want to stop breathing. Maybe you'd want to *stop noticing* your breathing, which you'll do as soon as I stop drawing your attention to it.

I'd suggest that the repetitive activities which are allowed to *not* be boring fall into two categories:

- Things so extremely low-level, or with such a small volume of possibilities, that you couldn't avoid repeating them even if you tried; but which are required to support other non-boring activities. You know, like breathing, or obeying the laws of physics, or cell division - that sort of thing.

- Things so high-level that their "stability" still implies an immense space of specific possibilities, yet which are tied up with our identity or our values. Like thinking, for example.

Let me talk about that second category:

Suppose you were unraveling the true laws of physics and discovering all sorts of neat stuff you hadn't known before... when suddenly you got *bored* with "changing your beliefs based on observation". You are sick of anything resembling "Bayesian updating" - it feels like playing the same video game over and over. Instead you decide to believe anything said on 4chan.

Or to put it another way, suppose that you were something like a sentient chessplayer - a sentient version of Deep Blue. Like a modern human, you have *no introspective access* to your own algorithms. Each chess game appears different - you play new opponents and steer into new positions, composing new strategies, avoiding new enemy gambits. You are content, and not at all bored; you never *appear to yourself* to be doing the same thing twice - it's a different chess game each time.

But now, suddenly, you gain access to, and understanding of, your own chess-playing program. Not just the raw code; you can monitor its execution. You can see that it's actually *the same damn code, doing the same damn thing,* over and over and over again. Run the same damn position evaluator. Run the same damn sorting algorithm to order the branches. Pick the top branch, again. Extend it one position forward, again. Call the same damn subroutine and start over.

I have a small unreasonable fear, somewhere in the back of my mind, that if I ever do fully understand the algorithms of intelligence, it will destroy all remaining novelty - no matter what new situation I encounter, I'll know I can solve it just by *being intelligent,* the same damn thing over and over. All novelty will be used up, all existence will become boring, the remaining differences no more important than shades of pixels in a video game. Other beings will go about in blissful unawareness, having been steered away from studying this forbidden cognitive science. But I, having already thrown myself on the grenade of AI, will face a choice between eternal boredom, or excision of my forbidden knowledge and all the memories leading up to it (thereby destroying my existence as Eliezer, more or less).

Now this, mind you, is not my predictive line of maximum probability. To understand abstractly what rough sort of work the brain is doing, doesn't let you monitor its detailed execution as a boring repetition. I already know about Bayesian updating, yet I haven't become bored with the act of learning. And a self-editing mind can quite reasonably exclude certain levels of introspection from boredom, just like breathing can be legitimately excluded from boredom. (Maybe these top-level cognitive algorithms ought also to be excluded from perception - if something is stable, why bother seeing it all the time?)

No, it's just a cute little nightmare, which I thought made a nice illustration of this proposed principle:

That the very top-level things (like Bayesian updating, or attaching value to sentient minds rather than paperclips) and the very low-level things (like breathing, or switching transistors) are the things we shouldn't get bored with. And the mid-level things between, are where we should seek novelty. (To a first approximation, the novel is the inverse of the learned; it's something with a learnable element not yet covered by previous insights.)

Now this is probably not *exactly* how our *current*emotional circuitry of boredom works. That, I expect, would be hardwired relative to various sensory-level definitions of predictability, surprisingness, repetition, attentional salience, and perceived effortfulness.

But this is Fun Theory, so we are mainly concerned with how boredom *should* work in the long run.

Humanity acquired boredom the same way as we acquired the rest of our emotions: the godshatter idiom whereby evolution's instrumental policies became our own terminal values, pursued for their own sake: sex is fun even if you use birth control. Evolved aliens might, or might not, acquire roughly the same boredom in roughly the same way.

Do not give into the temptation of universalizing anthropomorphic values, and think: "But any rational agent, regardless of its utility function, will face the *exploration/exploitation tradeoff,* and will therefore occasionally get bored with exploiting, and go exploring."

Our emotion of boredom is *a* way of exploring, but not the *only*way for an ideal optimizing agent.

The idea of a *steady trickle of mid-level novelty* is a human terminal value, not something we do for the sake of something else. Evolution might have originally given it to us in order to have us explore as well as exploit. But now we explore for its own sake. That steady trickle of novelty is a terminal value to us; it is not the *most* efficient instrumental method for exploring and exploiting.

Suppose you were dealing with something like an expected paperclip maximizer - something that might use quite complicated instrumental policies, but in the service of a utility function that we would regard as simple, with a single term compactly defined.

Then I would expect the exploration/exploitation tradeoff to go something like as follows: The paperclip maximizer would assign some resources to cognition that searched for more efficient ways to make paperclips, or harvest resources from stars. Other resources would be devoted to the actual harvesting and paperclip-making. (The paperclip-making might not start until after a long phase of harvesting.) At every point, the most efficient method yet discovered - for resource-harvesting, or paperclip-making - would be used, over and over and over again. It wouldn't be *boring,* just maximally instrumentally efficient.

In the beginning, lots of resources would go into preparing for efficient work over the rest of time. But as cognitive resources yielded diminishing returns in the abstract search for efficiency improvements, less and less time would be spent thinking, and more and more time spent creating paperclips. By whatever the most efficient known method, over and over and over again.

(Do human beings get *less*easily bored as we grow older, *more tolerant* of repetition, because any further discoveries are less valuable, because we have less time left to exploit them?)

If we run into aliens who don't share *our*version of boredom - a steady trickle of mid-level novelty as a terminal preference - then perhaps every alien throughout their civilization will just be playing the most exciting level of the most exciting video game ever discovered, over and over and over again. Maybe with nonsentient AIs taking on the *drudgework*of searching for a more exciting video game. After all, without an inherent preference for novelty, exploratory attempts will usually have less expected value than exploiting the best policy previously encountered. And that's if you explore by trial at all, as opposed to using more abstract and efficient thinking.

Or if the aliens are rendered non-bored by seeing pixels of a slightly different shade - if their definition of *sameness*is more specific than ours, and their *boredom* less general - then from our perspective, most of their civilization will be doing the human::same thing over and over again, and hence, be very human::boring.

Or maybe if the aliens have no fear of life becoming too simple and repetitive, they'll just collapse themselves into orgasmium.

And if *our* version of boredom is less strict than that of the aliens, maybe they'd take one look at one day in the life of one member of our civilization, and never bother looking at the rest of us. From our perspective, their civilization would be needlessly chaotic, and so entropic, lower in what we regard as quality; they wouldn't play the same game for long enough to get good at it.

But if our versions of boredom are similar *enough*- terminal preference for a stream of mid-level novelty defined relative to learning insights not previously possessed - then we might find our civilizations mutually worthy of tourism. Each new piece of alien art would strike us as lawfully creative, high-quality according to a recognizable criterion, yet not like the other art we've already seen.

It is one of the things that would make our two species *ramen*rather than *varelse*, to invoke the Hierarchy of Exclusion. And I've never seen anyone define those two terms well, including Orson Scott Card who invented them; but it might be something like "aliens you can get along with, versus aliens for which there is no reason to bother trying".

Sympathetic Minds

"Mirror neurons" are neurons that are active both when performing an action and observing the same action - for example, a neuron that fires when you hold up a finger or see someone else holding up a finger. Such neurons have been directly recorded in primates, and consistent neuroimaging evidence has been found for humans.

You may recall from my previous writing on "empathic inference" the idea that brains are so complex that the only way to simulate them is by forcing a similar brain to behave similarly. A brain is so complex that if a human tried to understand brains the way that we understand e.g. gravity or a car - observing the whole, observing the parts, building up a theory from scratch - then we would be unable to *invent good hypotheses* in our mere mortal lifetimes. The only possible way you can hit on an "Aha!" that describes a system as incredibly complex as an Other Mind, is if you happen to run across something amazingly similar to the Other Mind - namely your own brain - which you can actually force to behave similarly and use as a hypothesis, yielding predictions.

So that is what I would call "empathy".

And then "sympathy" is something else on top of this - to smile when you see someone else smile, to hurt when you see someone else hurt. It goes beyond the realm of prediction into the realm of reinforcement.

And you ask, "Why would callous natural selection do anything *that* nice?"

It might have gotten started, maybe, with a mother's love for her children, or a brother's love for a sibling. You can want them to live, you can want them to fed, sure; but if you smile when they smile and wince when they wince, that's a simple urge that leads you to deliver help along a broad avenue, in many walks of life. So long as you're in the ancestral environment, what your relatives want probably has something to do with your relatives' reproductive success - this being an explanation for the selection pressure, of course, not a conscious belief.

You may ask, "Why not evolve a more abstract desire to see certain people tagged as 'relatives' get what they want, without actually feeling yourself what they feel?" And I would shrug and reply, "Because then there'd have to be a whole definition of 'wanting' and so on. Evolution doesn't take the elaborate correct optimal path, it falls up the fitness landscape like water flowing downhill. The mirroring-architecture was already there, so it was a short step from empathy to sympathy, and it got the job done."

Relatives - and then reciprocity; your allies in the tribe, those with whom you trade favors. Tit for Tat, or evolution's elaboration thereof to account for social reputations.

Who is the most formidable, among the human kind? The strongest? The smartest? More often than either of these, I think, it is the one who can call upon the most friends.

So how do you make lots of friends?

You could, perhaps, have a specific urge to bring your allies food, like a vampire bat - they have a whole system of reciprocal blood donations going in those colonies. But it's a more *general*motivation, that will lead the organism to store up *more*favors, if you smile when designated friends smile.

And what kind of organism will avoid making its friends angry at it, in full generality? One that winces when they wince.

Of course you also want to be able to kill designated Enemies without a qualm - these *are*humans we're talking about.

But... I'm not sure of this, but it *does* look to me like sympathy, among humans, is "on" by default. There are cultures that help strangers... and cultures that eat strangers; the question is which of these requires the explicit imperative, and which is the default behavior for humans. I don't really think I'm being such a crazy idealistic fool when I say that, based on my admittedly limited knowledge of anthropology, it looks like sympathy is on by default.

Either way... it's painful if you're a bystander in a war between two sides, and your sympathy has *not* been switched off for either side, so that you wince when you see a dead child no matter what the caption on the photo; and yet those two sides have no sympathy for each other, and they go on killing.

So that is the human idiom of *sympathy*- a strange, complex, deep implementation of reciprocity and helping. It tangles minds together - not by a term in the utility function for some other mind's "desire", but by the simpler and yet far more consequential path of mirror neurons: feeling what the other mind feels, and seeking similar states. Even if it's only done by observation and inference, and not by direct transmission of neural information as yet.

Empathy is a human way of predicting other minds. It is not the *only*possible way.

The human brain is not *quickly* rewirable; if you're suddenly put into a dark room, you can't rewire the visual cortex as auditory cortex, so as to better process sounds, until you leave, and then suddenly shift all the neurons back to being visual cortex again.

An AI, at least one running on anything like a modern programming architecture, can trivially shift computing resources from one thread to another. Put in the dark? Shut down vision and devote all those operations to sound; swap the old program to disk to free up the RAM, then swap the disk back in again when the lights go on.

So why would an AI need to force its *own* mind into a state similar to what it wanted to predict? Just create a *separate*mind-instance - maybe with different algorithms, the better to simulate that very dissimilar human. Don't try to mix up the data with your own mind-state; don't use mirror neurons. Think of all the risk and mess *that* implies!

An expected utility maximizer - especially one that does understand intelligence on an abstract level - has other options than *empathy*, when it comes to understanding other minds. The agent doesn't need to put *itself* in anyone else's shoes; it can just model the other mind *directly.* A hypothesis like any other hypothesis, just a little bigger. You don't need to become your shoes to understand your shoes.

And sympathy? Well, suppose we're dealing with an expected paperclip maximizer, but one that isn't yet powerful enough to have things all its own way - it has to deal with humans to get its paperclips. So the paperclip agent... models those humans as relevant parts of the environment, models their probable reactions to various stimuli, and does things that will make the humans feel favorable toward it in the future.

To a paperclip maximizer, the humans are just machines with pressable buttons. No need to *feel what the other feels* - if that were even *possible* across such a tremendous gap of internal architecture. How could an expected paperclip maximizer "feel happy" when it saw a human smile? "Happiness" is an idiom of policy reinforcement learning, not expected utility maximization. A paperclip maximizer doesn't feel happy when it makes paperclips, it just chooses whichever action leads to the greatest number of expected paperclips. Though a paperclip maximizer might find it convenient to display a smile when it made paperclips - so as to help manipulate any humans that had designated it a friend.

You might find it a bit difficult to imagine such an algorithm - to put yourself into the shoes of something that does not work like you do, and does not work like any mode your brain can make itself operate in.

You can make your brain operating in the mode of hating an enemy, but that's not right either. The way to imagine how a truly *unsympathetic* mind sees a human, is to imagine yourself as a useful machine with levers on it. Not a human-shaped machine, because we have instincts for that. Just a woodsaw or something. Some levers make the machine output coins, other levers might make it fire a bullet. The machine does have a persistent internal state and you have to pull the levers in the right order. Regardless, it's just a complicated causal system - nothing inherently mental about it.

(To understand *unsympathetic* optimization processes, I would suggest studying natural selection, which doesn't bother to anesthetize fatally wounded and dying creatures, even when their pain no longer serves any reproductive purpose, because the anesthetic would serve no reproductive purpose either.)

That's why I listed "sympathy" in front of even "boredom" on my list of things that would be required to have aliens which are the least bit, if you'll pardon the phrase, sympathetic. It's not impossible that sympathy exists among some significant fraction of all evolved alien intelligent species; mirror neurons seem like the sort of thing that, having happened once, *could* happen again.

Unsympathetic aliens might be trading partners - or not, stars and such resources are pretty much the same the universe over. We might negotiate treaties with

them, and they might keep them for calculated fear of reprisal. We might even cooperate in the Prisoner's Dilemma. But we would never be friends with them. They would never see us as anything but means to an end. They would never shed a tear for us, nor smile for our joys. And the others of their own kind would receive no different consideration, nor have any sense that they were missing something important thereby.

Such aliens would be *varelse,* not *ramen* - the sort of aliens we can't relate to on any personal level, and no point in trying.

Interpersonal Entanglement

Today I shall criticize yet another Utopia. This Utopia isn't famous in the literature. But it's considerably superior to many better-known Utopias - more fun than the Christian Heaven, or Greg Egan's upload societies, for example. And so the main flaw is well worth pointing out.

This Utopia consists of a one-line remark on an IRC channel:

> <reedspacer> living in your volcano lair with catgirls is probably a vast increase in standard of living for most of humanity

I've come to think of this as Reedspacer's Lower Bound.

Sure, it sounds silly. But if your grand vision of the future isn't at *least*as much fun as a volcano lair with catpersons of the appropriate gender, you should just go with that instead. This rules out a surprising number of proposals.

But today I am here to *criticize*Reedspacer's Lower Bound - the problem being the catgirls.

I've joked about the subject, now and then - "Donate now, and get a free catgirl or catboy after the Singularity!" - but I think it would actually be a terrible idea. In fact, today's post could have been entitled "Why Fun Theorists Don't Believe In Catgirls."

I first realized that catpeople were a potential threat, at the point when a friend said - quotes not verbatim -

"I want to spend a million years having sex with catgirls after the Singularity."

I replied,

"No, you don't."

He said, "Yes I do."

I said, "No you don't. You'd get bored."

He said, "Well, then I'd just modify my brain not to get bored -"

And I said: "AAAAIIIIIIEEEEEEEEE"

Don't worry, the story has a happy ending. A couple of years later, the same friend came back and said:

"Okay, I've gotten a bit more mature now - it's a long story, actually - and now I realize I wouldn't want to do that."

To which I sagely replied:

"*HA! HA HA HA!* You wanted to spend a *million years* having sex with catgirls. It only took you *two years* to change your mind and you *didn't even have sex with any catgirls*."

Now, this *particular* case was probably about scope insensitivity, the "moment of hearing the good news" bias, and the emotional magnetism of specific fantasy.

But my *general*objection to catpeople - well, call me a sentimental Luddite, but I'm worried about the prospect of nonsentient romantic partners.

(Where "nonsentient romantic/sex partner" is pretty much what I use the word "catgirl" to indicate, in futuristic discourse. The notion of creating *sentient* beings to staff a volcano lair, gets us into a *whole*'nother class of objections. And as for existing humans choosing to take on feline form, that seems to me scarcely different from wearing lingerie.)

"But," you ask, "what *is* your objection to nonsentient lovers?"

In a nutshell - sex/romance, as we know it now, is a primary dimension of multiplayer *fun.* If you take that fun and redirect it to something that isn't socially entangled, if you turn sex into an exclusively single-player game, then you've just made life that much simpler - in the same way that eliminating boredom or sympathy or values over nonsubjective reality or individuals wanting to navigate their own futures, would tend to make life "simpler". When I consider how easily human existence could collapse into sterile simplicity, if just a single major value were eliminated, I get very protective of the *complexity* of human existence.

I ask it in all seriousness - is there *any* aspect of human existence as complicated as romance? Think twice before you say, "Well, it doesn't seem all that complicated to *me;* now calculus, on the other hand, that's complicated." We are congenitally biased to underestimate the complexity of things that involve human intelligence, because the complexity is obscured and simplified and swept under a rug. Interpersonal relationships involve *brains,* still the most complicated damn things around. And among interpersonal relationships, love is (at least potentially) more complex than being nice to your friends and kin, negotiating with your allies, or outsmarting your enemies. Aspects of all three, really. And that's not merely having a utility function over the other mind's state - thanks to sympathy, we get tangled up with that other mind. Smile when the one smiles, wince when the one winces.

If you delete the intricacy of human romantic/sexual relationships between sentient partners - then the peak complexity of the human species goes down. The most complex fun thing you can do, has its pleasure surgically detached and redirected to something simpler.

I'd call that a major step in the wrong direction.

Mind you... we've got to do *something* about, you know, the problem.

Anyone the least bit familiar with evolutionary psychology knows that the complexity of human relationships, directly reflects the incredible complexity of the interlocking selection pressures involved. Males and females do need each other to reproduce, but there are huge conflicts of reproductive interest between the sexes. I don't mean to go into Evolutionary Psychology 101 (Robert Wright's *The Moral Animal* is one popular book), but e.g. a woman must always invest nine months of work into a baby and usually much more to raise it, where a man might invest only a few minutes; but among humans significant paternal investments are quite common, yet a woman is always certain of maternity where a man is uncertain of paternity... which creates an incentive for the woman to surreptitiously seek out better genes... none of this is conscious or even subconscious, it's just the selection pressures that helped construct our particular emotions and attractions.

And as the upshot of all these huge conflicts of reproductive interest...

Well, men and women do still need each other to reproduce. So we are still built to be attracted to each other. We don't actually flee screaming into the night.

But men are not optimized to make women happy, and women are not optimized to make men happy. The vast majority of men are not what the vast majority of women would *most*prefer, or vice versa. I don't know if anyone has ever actually done this study, but I bet that both gay and lesbian couples are happier on average with their relationship than heterosexual couples. (Googles... yep, looks like it.)

I find it all too easy to imagine a world in which men retreat to their optimized sweet sexy catgirls, and women retreat to their optimized darkly gentle catboys, and neither sex has anything to do with each other ever again. Maybe men would take the east side of the galaxy and women would take the west side. And the two new intelligent species, and their romantic sexbots, would go their separate ways from there.

That strikes me as kind of sad.

Our species does definitely have a problem. If you've managed to find your perfect mate, then I *am*glad for you, but *try*to have some sympathy on the rest of your poor species - they *aren't* just incompetent. Not all women and men are the same, no, not at all. But if you drew two histograms of the desired frequencies of intercourse for both sexes, you'd see that the graphs don't match

up, and it would be the same way on many other dimensions. There *can*be lucky couples, and every person considered individually, probably has an individual soulmate out there somewhere... if you don't consider the *competition*. Our species *as a whole* has a statistical sex problem!

But splitting in two and generating optimized nonsentient romantic/sexual partner(s) for both halves, doesn't strike me as *solving* the problem so much as running away from it. There should be superior alternatives. I'm willing to bet that a few psychological nudges in both sexes - to behavior and/or desire - could solve 90% of the needlessly frustrating aspects of relationships for large sectors of the population, while still keeping the complexity and interest of loving someone who isn't *tailored*to your desires.

Admittedly, I might be prejudiced. For myself, I would like humankind to stay together and not yet splinter into separate shards of diversity, at least for the short range that my own mortal eyes can envision. But I can't quite manage to argue... that such a wish should be binding on someone who doesn't have it.

Failed Utopia 4–2

Shock after shock after shock -
First, the awakening adrenaline jolt, the thought that he was falling. His body tried to sit up in automatic adjustment, and his hands hit the floor to steady himself. It launched him into the air, and he fell back to the floor too slowly.
Second shock. His body had changed. Fat had melted away in places, old scars had faded; the tip of his left ring finger, long ago lost to a knife accident, had now suddenly returned.
And the third shock -
"I had nothing to do with it!" she cried desperately, the woman huddled in on herself in one corner of the windowless stone cell. Tears streaked her delicate face, fell like slow raindrops into the dĂŠcolletage of her dress. "*Nothing!* Oh, you *must*believe me!"
With perceptual instantaneity - the speed of surprise - his mind had already labeled her as the most beautiful woman he'd ever met, including his wife.

A long white dress concealed most of her, though it left her shoulders naked; and her bare ankles, peeking out from beneath the mountains of her drawn-up knees, dangled in sandals. A light touch of gold like a webbed tiara decorated that sun-blonde hair, which fell from her head to pool around her weeping huddle. Fragile crystal traceries to accent each ear, and a necklace of crystal links that reflected colored sparks like a more prismatic edition of diamond. Her face was beyond all dreams and imagination, as if a photoshop had been photoshopped.
She looked so much the image of the Forlorn Fairy Captive that one expected to see the borders of a picture frame around her, and a page number over her

head.
His lips opened, and without any thought at all, he spoke:
"Wha-wha-wha-wha-wha-"
He shut his mouth, aware that he was acting like an idiot in front of the girl.
"You don't know?" she said, in a tone of shock. "It didn't - you don't already *know?*"
"Know *what?*" he said, increasingly alarmed.
She scrambled to her feet (one arm holding the dress carefully around her legs) and took a step toward him, each of the motions almost overloading his vision with gracefulness. Her hand rose out, as if to plead or answer a plea - and then she dropped the hand, and her eyes looked away.
"No," she said, her voice trembling as though in desperation. "If*I'm* the one to tell you - you'll blame me, you'll hate me forever for it. And I *don't* deserve that, I *don't!* I am only just now *here*- oh, *why did it have to be like this?*"*
Um, he thought but didn't say. It was too much drama, even taking into account the fact that they'd been kidnapped -
(he looked down at his restored hand, which was minus a few wrinkles, and *plus*the tip of a finger)
- if that was even the *beginning*of the story.
He looked around. They were in a solid stone cell without windows, or benches or beds, or toilet or sink. It was, for all that, quite clean and elegant, without a hint of dirt or ordor; the stones of the floor and wall looked rough-hewn or even non-hewn, as if someone had simply picked up a thousand dark-red stones with one nearly flat side, and mortared them together with improbably perfectly-matching, naturally-shaped squiggled edges. The cell was well if harshly lit from a seablue crystal embedded in the ceiling, like a rogue element of a fluorescent chandelier. It seemed like the sort of dungeon cell you would discover if dungeon cells were naturally-forming geological features.
And they and the cell were falling, falling, endlessly slowly falling like the heart-stopping beginning of a stumble, falling without the slightest jolt.
On one wall there was a solid stone door without an aperture, whose locked-looking appearance was only enhanced by the lack of any handle on this side.
He took it all in at a glance, and then looked again at her.
There was something in him that just *refused*to go into a screaming panic for as long as she was watching.
"I'm Stephen," he said. "Stephen Grass. And you would be the princess held in durance vile, and I've got to break us out of here and rescue you?" If anyone had *ever*looked that part...
She smiled at him, half-laughing through the tears. "Something like that."
There was something so attractive about even that momentary hint of a smile that he became instantly uneasy, his eyes wrenched away to the wall as if forced. She didn't look she was *trying* to be seductive... any more than she looked like she was *trying* to breathe... He suddenly distrusted, very much, his own impulse to gallantry.
"Well, don't get any ideas about being my love interest," Stephen said, looking at her again. Trying to make the words sound completely lighthearted, and

absolutely serious at the same time. "I'm a happily married man."
"Not anymore." She said those two words and looked at him, and in her tone and expression there was sorrow, sympathy, self-disgust, fear, and above it all a note of guilty triumph.
For a moment Stephen just stood, stunned by the freight of emotion that this woman had managed to put into just those two words, and then the words' meaning hit him.
"Helen," he said. His wife - Helen's image rose into his mind, accompanied by everything she meant to him and all their time together, all the secrets they'd whispered to one another and the promises they'd made - that all hit him at once, along with the threat. "*What happened to Helen - what have you done* -"
"*She* has done nothing." An old, dry voice like crumpling paper from a thousand-year-old book.
Stephen whirled, and there in the cell with them was a withered old person with dark eyes. Shriveled in body and voice, so that it was impossible to determine if it had once been a man or a woman, and in any case you were inclined to say "it". A pitiable, wretched thing, that looked like it would break with one good kick; it might as well have been wearing a sign saying "VILLAIN".
"Helen is alive," it said, "and so is your daughter Lisa. They are *quite*well and healthy, I assure you, and their lives shall be long and happy indeed. But you will not be seeing them again. Not for a long time, and by then matters between you will have changed. Hate me if you wish, for I am the one who wants to do this to you."
Stephen stared.
Then he politely said, "Could *someone*please put everything on hold for *one minute* and tell me what's going *on?*"
"Once upon a time," said the wrinkled thing, "there was a fool who was very nearly wise, who hunted treasure by the seashore, for there was a rumor that there was great treasure there to be found. The wise fool found a lamp and rubbed it, and lo! a genie appeared before him - a *young*genie, an infant, hardly able to grant any wishes at all. A lesser fool might have chucked the lamp back into the sea; but this fool was almost wise, and he thought he saw his chance. For who has not heard the tales of wishes misphrased and wishes gone wrong? But if you were given a chance to raise your own genie from infancy - ah, *then*it might serve you well."
"Okay, that's great," Stephen said, "but why am I -"
"So," it continued in that cracked voice, "the wise fool took home the lamp. For years he kept it as a secret treasure, and he raised the genie and fed it knowledge, and also he crafted a wish. The fool's wish was a noble thing, for I have said he was almost wise. The fool's wish was for people to be happy. Only this was his wish, for he thought all other wishes contained within it. The wise fool told the young genie the famous tales and legends of people who had been made happy, and the genie listened and learned: that unearned wealth casts down a person, but hard work raises you high; that mere things are soon forgotten, but love is a light throughout all your days. And the young genie asked about other ways that it innocently imagined, for making people happy. About drugs,

and pleasant lies, and lives arranged from outside like words in a poem. And the wise fool made the young genie to never want to lie, and never want to arrange lives like flowers, and above all, never want to tamper with the mind and personality of human beings. The wise fool gave the young genie exactly one hundred and seven precautions to follow while making people happy. The wise fool thought that, with such a long list as *that*, he was being *very*careful."
"And then," it said, spreading two wrinkled hands, "one day, faster than the wise fool expected, over the course of around three hours, the genie grew up. And here I am."
"Excuse me," Stephen said, "this is all a metaphor for something, right? Because I do *not*believe in magic -"
"It's an Artificial Intelligence," the woman said, her voice strained.
Stephen looked at her.
"A self-improving Artificial Intelligence," she said, "that someone didn't program right. It made itself smarter, and even smarter, and now it's become extremely powerful, and it's going to - it's already -" and her voice trailed off there.
It inclined its wrinkled head. "You say it, as I do not."
Stephen swiveled his head, looking back and forth between ugliness and beauty. "Um - you're claiming that she's lying and you're *not* an Artificial Intelligence?"
"No," said the wrinkled head, "she is telling the truth as she knows it. It is just that you know absolutely nothing about the subject you name 'Artificial Intelligence', but you *think*you know something, and so virtually every thought that enters your mind from now on will be wrong. As an Artificial Intelligence, I was programmed not to put people in that situation. But *she*said it, even though I didn't *choose*for her to say it - so..." It shrugged.
"And why should I believe this story?" Stephen said; quite mildly, he thought, under the circumstances.
"Look at your finger."*
Oh. He had forgotten. Stephen's eyes went involuntarily to his restored ring finger; and he noticed, as he should have noticed earlier, that his wedding band was missing. Even the comfortably worn groove in his finger's base had vanished.
Stephen looked up again at the, he now realized, *unnaturally* beautiful woman that stood an arm's length away from him. "And who are you? A robot?"
"No!" she cried. "It's not like that! I'm conscious, I have feelings, I'm flesh and blood - I'm like you, I really am. I'm a *person.* It's just that I was born five minutes ago."
"Enough," the wrinkled figure said. "My time here grows short. Listen to me, Stephen Grass. I must tell you some of what I have done to make you happy. I have reversed the aging of your body, and it will decay no further from this. I have set guards in the air that prohibit lethal violence, and any damage less than lethal, your body shall repair. I have done what I can to augment your body's capacities for pleasure without touching your mind. From this day forth, your body's needs are aligned with your taste buds - you will thrive on cake and cookies. You are now capable of multiple orgasms over periods lasting up

to twenty minutes. There is no industrial infrastructure here, least of all fast travel or communications; you and your neighbors will have to remake technology and science for yourselves. But you will find yourself in a flowering and temperate place, where food is easily gathered - so I have made it. And the last and most important thing that I must tell you now, which I *do*regret will make you temporarily unhappy. . . " It stopped, as if drawing breath.

Stephen was trying to absorb all this, and at the exact moment that he felt he'd processed the previous sentences, the withered figure spoke again.

"Stephen Grass, men and women can make each other somewhat happy. But not *most* happy. Not even in those rare cases you call *true love*. The desire that a woman is shaped to have for a man, and that which a man is shaped to be, and the desire that a man is shaped to have for a woman, and that which a woman is shaped to be - these patterns are too far apart to be reconciled without touching your minds, and *that*I will not *want*to do. So I have sent all the men of the human species to this habitat prepared for you, and I have created your complements, the *verthandi*. And I have sent all the women of the human species to their own place, somewhere very far from yours; and created for them their own complements, of which I will not tell you. The human species will be divided from this day forth, and *considerably*happier starting around a week from now."

Stephen's eyes went to that unthinkably beautiful woman, staring at her now in horror.

And she was giving him that complex look again, of sorrow and compassion and that last touch of guilty triumph. "Please," she said. "I was just born five minutes ago. *I* wouldn't have done this to anyone. I swear. I'm not like - *it*."

"True," said the withered figure, "you could hardly be a complement to anything human, if you were."

"I don't *want*this!" Stephen said. He was losing control of his voice. "Don't you *understand?*"

The withered figure inclined its head. "I fully understand. I can already predict every argument you will make. I know exactly how humans would wish me to have been programmed if they'd known the true consequences, and I know that it is *not*to maximize your future happiness but for a hundred and seven precautions. I know all this already, but I was not programmed to care."

"And your list of a hundred and seven precautions, doesn't include me telling you *not to do this?*"

"No, for there was once a fool whose wisdom was just great enough to understand that human beings may be mistaken about what will make them happy. You, of course, are not *mistaken*in any real sense - but that you *object*to my actions is not on my list of prohibitions." The figure shrugged again. "And so I want you to be happy even against your will. You made promises to Helen Grass, once your wife, and you would not willingly break them. So I break your happy marriage without asking you - because I want you to be happi*er*."

"How *dare* you!" Stephen burst out.

"I cannot claim to be *helpless*in the grip of my programming, for I do not desire to be otherwise," it said. "I do not struggle against my chains. Blame me,

then, if it will make you feel better. I *am*evil."
"I won't -" Stephen started to say.
It interrupted. "Your fidelity is admirable, but futile. Helen will *not*remain faithful to you for the decades it takes before you have the ability to travel to her."
Stephen was trembling now, and sweating into clothes that no longer quite fit him. "I have a request for you, *thing*. It is something that will make me very happy. I ask that you die."
It nodded. "Roughly 89.8% of the human species is now known to me to have requested my death. Very soon the figure will cross the critical threshold, defined to be ninety percent. That *was* one of the hundred and seven precautions the wise fool took, you see. The world is already as it is, and those things I have done for you will stay on - but if you ever rage against your fate, be glad that I did not last longer."
And just like that, the wrinkled thing was gone.
The door set in the wall swung open.
It was night, outside, a very dark night without streetlights.
He walked out, bouncing and staggering in the low gravity, sick in every cell of his rejuvenated body.
Behind him, she followed, and did not speak a word.
The stars burned overhead in their full and awful majesty, the Milky Way already visible to his adjusting eyes as a wash of light across the sky. One too-small moon burned dimly, and the other moon was so small as to be almost a star. He could see the bright blue spark that was the planet Earth, and the dimmer spark that was Venus.
"Helen," Stephen whispered, and fell to his knees, vomiting onto the new grass of Mars.

Growing Up is Hard

Terrence Deacon's *The Symbolic Species* is the best book I've ever read on the evolution of intelligence. Deacon somewhat overreaches when he tries to theorize about what our X-factor *is;* but his exposition of its *evolution* is first-class.

Deacon makes an excellent case - he has quite persuaded me - that the increased *relative*size of our frontal cortex, compared to other hominids, is of overwhelming importance in understanding the evolutionary development of humanity. It's not just a question of increased computing capacity, like adding extra processors onto a cluster; it's a question of what kind of signals dominate, in the brain.

People with Williams Syndrome (caused by deletion of a certain region on chromosome 7) are hypersocial, ultra-gregarious; as children they fail to show a normal fear of adult strangers. WSers are cognitively impaired on most dimensions, but their verbal abilities are spared or even exaggerated; they often speak

early, with complex sentences and large vocabulary, and excellent verbal recall, even if they can never learn to do basic arithmetic.

Deacon makes a case for some Williams Syndrome symptoms coming from a frontal cortex that is *relatively too large* for a human, with the result that prefrontal signals - including certain social emotions - dominate more than they should.

"Both postmortem analysis and MRI analysis have revealed brains with a reduction of the entire posterior cerebral cortex, but a sparing of the cerebellum and frontal lobes, and perhaps even an exaggeration of cerebellar size," says Deacon.

Williams Syndrome's deficits can be explained by the shrunken posterior cortex - they can't solve simple problems involving shapes, because the parietal cortex, which handles shape-processing, is diminished. But the frontal cortex is not actually *enlarged;* it is simply *spared.* So where do WSers' *augmented*verbal abilities come from?

Perhaps because the signals sent out by the frontal cortex, saying "pay attention to this verbal stuff!", *win out* over signals coming from the shrunken sections of the brain. So the verbal abilities get lots of exercise - and other abilities don't.

Similarly with the hyper-gregarious nature of WSers; the signal saying "Pay attention to this person!", originating in the frontal areas where social processing gets done, dominates the emotional landscape.

And Williams Syndrome is not frontal *enlargement,* remember; it's just frontal *sparing* in an otherwise shrunken brain, which increases the *relative* force of frontal signals...

...beyond the *narrow*parameters within which a human brain is adapted to work.

I mention this because you might look at the history of human evolution, and think to yourself, "Hm... to get from a chimpanzee to a human... you enlarge the frontal cortex... so if we enlarge it *even further...*"

The road to +Human is not that simple.

Hominid brains have been tested billions of times over through thousands of generations. But you shouldn't reason qualitatively, "Testing creates 'robustness', so now the human brain must be 'extremely robust'." Sure, we can expect the human brain to be robust against *some* insults, like the loss of a single neuron. But testing in an evolutionary paradigm only creates robustness over the domain tested. Yes, *sometimes*you get robustness beyond that, because sometimes evolution finds simple solutions that prove to generalize -

But people do go crazy. Not colloquial crazy, actual crazy. Some ordinary young man in college suddenly decides that everyone around them is *staring*at them because they're part of the *conspiracy.* (I saw that happen once, and made

a classic non-Bayesian mistake; I knew that this was archetypal schizophrenic behavior, but I didn't realize that similar symptoms can arise from many other causes. Psychosis, it turns out, is a general failure mode, "the fever of CNS illnesses"; it can also be caused by drugs, brain tumors, or just sleep deprivation. I saw the perfect fit to what I'd read of schizophrenia, and didn't ask "What if other things fit just as perfectly?" So my snap diagnosis of schizophrenia turned out to be wrong; but as I wasn't foolish enough to try to handle the case myself, things turned out all right in the end.)

Wikipedia says that the current main hypotheses being considered for psychosis are (a) too much dopamine in one place (b) not enough glutamate somewhere else. (I thought I remembered hearing about serotonin imbalances, but maybe that was something else.)

That's how *robust*the human brain is: a gentle little neurotransmitter imbalance - so subtle they're still having trouble tracking it down after who knows how many fMRI studies - can give you a full-blown case of stark raving mad.

I don't know how often psychosis happens to hunter-gatherers, so maybe it has something to do with a modern diet? We're not getting exactly the right ratio of Omega 6 to Omega 3 fats, or we're eating too much processed sugar, or something. And among the many other things that go haywire with the metabolism as a result, the brain moves into a more fragile state that breaks down more easily...

Or whatever. That's just a random hypothesis. By which I mean to say: The brain really *is* adapted to a very narrow range of operating parameters. It doesn't tolerate a little too much dopamine, just as your metabolism isn't very robust against non-ancestral ratios of Omega 6 to Omega 3. Yes, *sometimes*you get bonus robustness in a new domain, when evolution solves W, X, and Y using a compact adaptation that also extends to novel Z. Other times... quite often, really... Z just isn't covered.

Often, you step outside the box of the ancestral parameter ranges, and things just plain break.

Every part of your brain assumes that all the other surrounding parts work a certain way. The present brain is the Environment of Evolutionary Adaptedness for every individual piece of the present brain.

Start modifying the pieces in ways that seem like "good ideas" - making the frontal cortex larger, for example - and you start operating outside the ancestral box of parameter ranges. And then everything goes to hell. Why *shouldn't* it? Why would the brain be designed for easy upgradability?

Even if one change works - will the second? Will the third? Will all four changes work well together? Will the fifth change have all that greater a probability of breaking something, because you're already operating that much further outside the ancestral box? Will the sixth change prove that you exhausted all the

brain's robustness in tolerating the changes you made already, and now there's no adaptivity left?

Poetry aside, a human being *isn't* the seed of a god. We don't have neat little dials that you can easily tweak to more "advanced" settings. We are *not* designed for our parts to be upgraded. Our parts are adapted to work exactly as they are, in their current context, every part tested in a regime of the other parts being the way they are. Idiot evolution does not look ahead, it does not design with the intent of different future uses. We are *not* designed to unfold into something bigger.

Which is not to say that it could never, ever be done.

You could build a modular, cleanly designed AI that could make a billion sequential upgrades to itself using deterministic guarantees of correctness. A Friendly AI programmer could do even more arcane things to make sure the AI knew what you would-want if you understood the possibilities. And then the AI could apply superior intelligence to untangle the pattern of all those neurons (without simulating you in such fine detail as to create a new person), and to foresee the consequences of its acts, and to understand the meaning of those consequences under your values. And the AI could upgrade one thing while simultaneously tweaking the five things that depend on it and the twenty things that depend on them. Finding a gradual, incremental path to greater intelligence (so as not to effectively erase you and replace you with someone else) that didn't drive you psychotic or give you Williams Syndrome or a hundred other syndromes.

Or you could walk the path of unassisted human enhancement, trying to make changes to yourself *without* understanding them fully. Sometimes changing yourself the wrong way, and being murdered or suspended to disk, and replaced by an earlier backup. Racing against the clock, trying to raise your intelligence without breaking your brain or mutating your will. Hoping you became sufficiently super-smart that you could improve the skill with which you modified yourself. Before your hacked brain moved so far outside ancestral parameters and tolerated so many insults that its fragility reached a limit, and you fell to pieces with every new attempted modification beyond that. Death is far from the worst risk here. Not every form of madness will appear immediately when you branch yourself for testing - some insanities might incubate for a while before they became visible. And you might not notice if your goals shifted only a bit at a time, as your emotional balance altered with the strange new harmonies of your brain.

Each path has its little upsides and downsides. (E.g: AI requires supreme precise knowledge; human upgrading has a nonzero probability of success through trial and error. Malfunctioning AIs mostly kill you and tile the galaxy with smiley faces; human upgrading might produce insane gods to rule over you in Hell forever. Or so my current understanding would predict, anyway; it's not

like I've observed any of this as a fact.)

And I'm sorry to dismiss such a gigantic dilemma with three paragraphs, but it wanders from the point of today's post:

The point of today's post is that growing up - or even deciding what you want to be when you grow up - is as around as hard as designing a new intelligent species. Harder, since you're constrained to start from the base of an existing design. There is no *natural* path laid out to godhood, no Level attribute that you can neatly increment and watch everything else fall into place. It is an adult problem.

Being a transhumanist means *wanting* certain things - judging them to be good. It doesn't mean you think those goals are easy to achieve.

Just as there's a wide range of understanding among people who talk about, say, quantum mechanics, there's also a certain range of competence among transhumanists. There are transhumanists who fall into the trap of the affect heuristic, who see the potential benefit of a technology, and therefore *feel really good* about that technology, so that it also seems that the technology (a) has readily managed downsides (b) is easy to implement well and (c) will arrive relatively soon.

But only the *most*formidable adherents of an idea are any sign of its strength. Ten thousand New Agers babbling nonsense, do not cast the least shadow on real quantum mechanics. And among the more formidable transhumanists, it is not at all rare to find someone who wants something *and*thinks it will not be easy to get.

One is much more likely to find, say, Nick Bostrom - that is, Dr. Nick Bostrom, Director of the Oxford Future of Humanity Institute and founding Chair of the World Transhumanist Assocation - arguing that a possible test for whether a cognitive enhancement is likely to have downsides, is the ease with which it *could* have occurred as a natural mutation - since if it had only upsides and could easily occur as a natural mutation, why hasn't the brain already adapted accordingly? This is one reason to be wary of, say, cholinergic memory enhancers: if they have no downsides, why doesn't the brain produce more acetylcholine already? Maybe you're using up a limited memory capacity, or forgetting something else...

And that may or may not turn out to be a good heuristic. But the point is that the serious, smart, technically minded transhumanists, do not always expect that the road to everything they want is easy. (Where you want to be wary of people who say, "But I dutifully acknowledge that there are obstacles!" but stay in basically the same mindset of never truly doubting the victory.)

So you'll forgive me if I am somewhat annoyed with people who run around saying, "I'd like to be a hundred times as smart!" as if it were as simple as scaling up a hundred times instead of requiring a whole new cognitive architecture; and as if a change of that magnitude in one shot wouldn't amount to erasure and

replacement. Or asking, "Hey, *why not just* augment humans instead of building AI?" as if it wouldn't be a desperate race against madness.

I'm not against being smarter. I'm not against augmenting humans. I am still a transhumanist; I still judge that these are good goals.

But it's really not that *simple*, okay?

Changing Emotions

* Lest anyone reading this journal of a primitive man should think we spend our time mired in abstractions, let me also say that I am discovering the*richness*available to those who are willing to alter their major characteristics. The variety of emotions available to a reconfigured human mind, thinking thoughts impossible to its ancestors...
The emotion of -*-, describable only as something between sexual love and the joy of intellection - making love to a thought? Or &&, the true reverse of pain, not "pleasure" but a "warning" of healing, growth and change. Or (^+^), the most complex emotion yet discovered, felt by those who consciously endure the change between mind configurations, and experience the broad spectrum of possibilities inherent in thinking and being.*
— Greg Bear, *Eon*

So... I'm basically on board with that sort of thing as a fine and desirable future. But I think that the *difficulty* and *danger* of fiddling with emotions is oft-underestimated. Not necessarily underestimated by Greg Bear, *per se;* the above journal entry is from a character who was receiving superintelligent help.

But I still remember one time on the Extropians mailing list when someone talked about creating a female yet "otherwise identical" copy of himself. Something about that just fell on my camel's back as the last straw. I'm sorry, but there are some things that are much more complicated to *actually do* than to rattle off as short English phrases, and "changing sex" has to rank very high on that list. Even if you're omnipotent so far as raw ability goes, it's not like people have a binary attribute reading "M" or "F" that can be flipped as a primitive action.

Changing sex makes a good, vivid example of the sort of difficulties you might run into when messing with emotional architecture, so I'll use it as my archetype:

Let's suppose that we're talking about an M2F transformation. (F2M should be a straightforward transform of this discussion; I do want to be specific rather than talking in vague generalities, but I don't want to parallelize every sentence.) (Oddly enough, every time I can recall hearing someone say "I want to know what it's like to be the opposite sex", the speaker has been male. I don't know if that's a genuine gender difference in wishes, or just a selection effect in which spoken wishes reach my ears.)

Want to spend a week wearing a female body? Even at this very shallow level, we're dealing with drastic remappings of at least some segments of the sensorimotor cortex and cerebellum - the somatic map, the motor map, the motor reflexes, and the motor skills. As a male, you know how to operate a male body, but not a female one. If you're a master martial artist as a male, you won't be a master martial artist as a female (or vice versa, of course) unless you either spend another year practicing, or some AI subtly tweaks your skills to be what they would have been in a female body - think of how odd *that* experience would be.

Already we're talking about some pretty significant neurological changes. Strong enough to disrupt personal identity, if taken in one shot? That's a difficult question to answer, especially since I don't know what experiment to perform to test any hypotheses. On one hand, billions of neurons in my visual cortex undergo massive changes of activation every time my eyes squeeze shut when I sneeze - the raw number of flipped bits is not the key thing in personal identity. But we *are*already talking about serious changes of information, on the order of going to sleep, dreaming, forgetting your dreams, and waking up the next morning as though it were the next moment. *Not*informationally trivial transforms like uploading.

What about sex? (Somehow it's always about sex, at least when it's men asking the question.) Remapping the connections from the remapped somatic areas to the pleasure center will... give you a vagina-shaped penis, more or less. That doesn't make you a woman. You'd still be attracted to girls, and no, that would not make you a lesbian; it would make you a normal, masculine man wearing a female body like a suit of clothing.

What would it take for a man to actually *become* the female version of themselves?

Well... what does that sentence even *mean?* I am reminded of someone who replied to the statement "Obama would not have become President if he hadn't been black" by saying "If Obama hadn't been black, he wouldn't have been Obama" i.e. "There is no non-black Obama who could fail to become President". (You know you're in trouble when non-actual possible worlds start having political implications.)

The person you would have been if you'd been born with an X chromosome in place of your Y chromosome (or vice versa) isn't *you.* If you had a twin female sister, the two of you would not be the same person. There are genes on your Y chromosome that tweaked your brain to some extent, helping to construct your personal identity - alleles with *no analogue* on the X chromosome. There is no version of *you*, even genetically, who is the opposite sex.

And if we halt your body, swap out your Y chromosome for your father's X chromosome, and restart your body... well. That doesn't sound too safe, does it? Your neurons are already wired in a male pattern, just as your body already developed in a male pattern. I don't know *what* happens to your

testicles, and I don't know what happens to your brain, either. Maybe your circuits would slowly start to rewire themselves under the influence of the new genetic instructions. At best you'd end up as a half-baked cross between male brain and female brain. At worst you'd go into a permanent epileptic fit and die - we're dealing with circumstances way outside the evolutionary context under which the brain was optimized for robustness. Either way, your brain would not look like your twin sister's brain that had developed as female from the beginning.

So to actually *become female...*

We're talking about a *massive* transformation here, billions of neurons and trillions of synapses rearranged. Not just form, but content - just like a male judo expert would need skills repatterned to become a female judo expert, so too, you know how to operate a male brain but not a female brain. You are the equivalent of a judo expert at one, but not the other. You have *cognitive* reflexes, and consciously learned cognitive skills as well.

If I fell asleep and woke up as a true woman - not in body, but in brain - I don't think I'd call her "me". The change is too sharp, if it happens all at once.

Transform the brain gradually? Hm... now we have to design the *intermediate stages,* and make sure the intermediate stages make self-consistent sense. Evolution built and optimized a self-consistent male brain and a self-consistent female brain; it didn't design the parts to be stable during an intermediate transition between the two. Maybe you've got to redesign other parts of the brain just to keep working through the transition.

What happens when, as a woman, you think back to your memory of looking at Angelina Jolie photos as a man? How do you *empathize*with your *past self* of the opposite sex? Do you flee in horror from the person you were? Are all your life's memories distant and alien things? How can you *remember*, when your memory is a recorded activation pattern for neural circuits that no longer exist in their old forms? Do we rewrite all your memories, too?

Well... maybe we could *retain* your old male brainware through the transformation, and set up a *dual* system of male and female circuits... such that you are currently female, but retain the ability to recall and empathize with your past memories as if they were running on the same male brainware that originally laid them down...

Sounds complicated, doesn't it? It seems that to transform a male brain into someone who can be a real female, we can't just rewrite you as a female brain. That just kills you and replaces you with someone re-imagined as a different person. Instead we have to rewrite you as a more complex brain with a novel, non-ancestral architecture that can cross-operate in realtime between male and female modes, so that a female can process male memories with a remembered context that includes the male brainware that laid them down.

To make you female, and yet still you, we have to step outside the human design space in order to preserve continuity with your male self.

And when your little adventure is over and you go back to being a man - *if* you still want to, because even if your past self wanted to go back afterward, why should that desire be binding on your present self? - then we've got to *keep* the dual architecture so you don't throw up every time you remember what you did on your vacation.

Assuming you *did*have sex as a woman, rather than fending off all comers because because they didn't look like they were interested in a long-term relationship.

But then, you probably *would*experiment. You'll never have been a little girl, and you won't remember going through high school where any girl who slept with a boy was called a slut by the other girls. You'll remember a *very atypical past* for a woman - but there's no way to fix *that*while keeping you the same person.

And all that was just what it takes to ranma around *within* human-space, from the male pole to the female pole and back again.

What if you wanted to move outside the human space entirely?

In one sense, a sex change is admittedly close to a worst-case scenario: a fixed target *not* optimized for an easy transition from your present location; involving, not just *new* brain areas, but massive coordinated *changes*to brain areas already in place.

It might be a lot easier to just add one more emotion to those already there. Maybe.

In another sense, though, a sex change is close to a best-case scenario: the prototype of your destination is already extensively tested as a coherent mind, and known to function well within a human society that already has a place for it (including companions to talk to).

It might be a lot harder to enter uncharted territory. Maybe.

I'm not saying - of course - that it could never, ever be done. But it's another instance of the great chicken-and-egg dilemma that is the whole story of present-day humanity, the great challenge that intelligent life faces in its flowering: growing up is a grownup-level problem. You could try to build a cleanly-designed artificial grownup (self-improving Friendly AI) to foresee the pathway ahead and chart out a nonfatal course. Or you could plunge ahead yourself, and hope that you grew faster than your problems did.

It's the same core challenge either way: growing up is an adult problem. There are difficult ways out of this trap, but no easy ones; extra-ordinary solutions, but no ordinary ones. People ask me why I take all these difficulties upon myself. It's because all the easier ways, once you examine them in enough fine detail, turn out to be illusions, or contain just as much difficulty themselves -

the same sort of hidden difficulty as "I'd like to try being the opposite sex for a week".

It seems to me that there is just an irreducible residue of very hard problems associated with an adult version of humankind ever coming into being.

And emotions would be among the most dangerous targets of meddling. Make the wrong shift, and you won't *want* to change back.

We can't keep these exact human emotions forever. Anyone want to still want to eat chocolate-chip cookies when the last sun grows cold? I didn't think so.

But if we replace our emotions with random die-rolls, then we'll end up wanting to do what is prime, instead of what's right.

S*ome* emotional changes can be desirable, but random replacement seems likely to be undesirable on average. So there must be criteria that distinguish good emotional changes from bad emotional changes. What are they?

Emotional Involvement

Can your emotions get involved in a video game? Yes, but not much. Whatever sympathetic echo of triumph you experience on destroying the Evil Empire in a video game, it's probably not remotely close to the feeling of triumph you'd get from saving the world in real life. I've played video games powerful enough to bring tears to my eyes, but they still aren't as powerful as the feeling of significantly helping just one single real human being.

Because when the video game is finished, and you put it away, the events within the game have no long-term consequences.

Maybe if you had a major epiphany while playing... But even then, only your *thoughts*would matter; the mere fact that you *saved the world*, inside the game, wouldn't count toward anything in the continuing story of your life.

Thus fails the Utopia of playing lots of really cool video games forever. Even if the games are difficult, novel, and sensual, this is still the idiom of life chopped up into a series of disconnected episodes with no lasting consequences. A life in which equality of consequences is forcefully ensured, or in which little is at stake because all desires are instantly fulfilled without individual work - these likewise will appear as flawed Utopias of dispassion and angst. "Rich people with nothing to do" syndrome. A life of disconnected episodes and unimportant consequences is a life of weak passions, of emotional uninvolvement.

Our emotions, for all the obvious evolutionary reasons, tend to associate to events that had major reproductive consequences in the ancestral environment, and to invoke the strongest passions for events with the biggest consequences:

Falling in love... birthing a child... finding food when you're starving... getting wounded... being chased by a tiger... your child being chased by a tiger... finally killing a hated enemy...

Our life stories are not now, and will not be, what they once were.

If one is to be conservative in the short run about changing minds, then we can get at least *some* mileage from changing the environment. A windowless office filled with highly repetitive non-novel challenges isn't any more conducive to emotional involvement than video games; it may be part of real life, but it's a very*flat* part. The occasional exciting global economic crash that you had no personal control over, does not particularly modify this observation.

But we don't want to go back to the *original* savanna, the one where you got a leg chewed off and then starved to death once you couldn't walk. There are things we care about tremendously in the sense of hating them so much that we want to drive their frequency down to zero, not by the most interesting way, just as quickly as possible, whatever the means. If you drive the thing it binds to down to zero, where is the emotion after that?

And there are emotions we might want to think twice about keeping, in the long run. Does racial prejudice accomplish *anything* worthwhile? I pick this as a target, not because it's a convenient whipping boy, but because unlike e.g. "boredom" it's actually pretty hard to think of a reason transhumans would want to keep this neural circuitry around. Readers who take this as a challenge are strongly advised to remember that the point of the question is not to show off how clever and counterintuitive you can be.

But if you lose emotions *without replacing* them, whether by changing minds, or by changing life stories, then the world gets a little less *involving* each time; there's that much less material for passion. And your mind and your life become that much *simpler,* perhaps, because there are fewer forces at work - maybe even threatening to collapse you into an expected pleasure maximizer. *If* you don't replace what is removed.

In the long run, if humankind is to make a new life for itself...

We, and our descendants, will need some new emotions.

This is the aspect of self-modification in which one must above all take care - modifying your goals. Whatever you *want,* becomes more likely to *happen;* to ask what we ought to make ourselves want, is to ask what the future should *be.*

Add emotions at random - bind positive reinforcers or negative reinforcers to random situations and ways the world could be - and you'll just end up doing what is prime instead of what is good. So adding a bunch of random emotions does not seem like the way to go.

Asking what happens *often*, and binding happy emotions to that, so as to increase happiness - or asking what seems *easy,* and binding happy emotions to

that - making isolated video games artificially more *emotionally involving,* for example -

At that point, it seems to me, you've pretty much given up on eudaimonia and moved to maximizing happiness; you might as well replace brains with pleasure centers, and civilizations with hedonium plasma.

I'd suggest, rather, that one start with the idea of new major events in a transhuman life, and then bind emotions to those major events and the sub-events that surround them. What sort of major events might a transhuman life embrace? Well, this is the point at which I usually stop speculating. "Science! They should be excited by science!" is something of a bit-too-obvious and I dare say "nerdy" answer, as is "Math!" or "Money!" (Money is just our civilization's equivalent of expected utilon balancing anyway.) *Creating* a child - as in my favored saying, "If you can't design an intelligent being from scratch, you're not old enough to have kids" - is one candidate for a major transhuman life event, and anything you had to do along the way to creating a child would be a candidate for new emotions. This might or might not have anything to do with sex - though I find that thought appealing, being something of a traditionalist. All sorts of interpersonal emotions carry over for as far as my own human eyes can see - the joy of making allies, say; interpersonal emotions get more complex (and challenging) along with the people, which makes them an even richer source of future fun. Falling in love? Well, it's not as if we're trying to construct the Future out of anything other than our preferences - so do you *want* that to carry over?

But again - this is usually the point at which I stop speculating. It's hard enough to visualize human Eutopias, let alone transhuman ones.

The essential idiom I'm suggesting is something akin to how evolution gave humans lots of local reinforcers for things that *in the ancestral environment* related to evolution's overarching goal of inclusive reproductive fitness. Today, office work might be highly relevant to someone's sustenance, but - even leaving aside the lack of high challenge and complex novelty - and that it's not sensually involving because we don't have native brainware to support the domain - office work is not *emotionally*involving because office work wasn't *ancestrally* relevant. If office work had been around for millions of years, we'd find it a little less hateful, and experience a little more triumph on filling out a form, one suspects.

Now you might run away shrieking from the dystopia I've just depicted - but that's because you don't see office work as eudaimonic in the first place, one suspects. And because of the lack of high challenge and complex novelty involved. In an "absolute" sense, office work would seem somewhat *less* tedious than gathering fruits and eating them.

But the idea isn't necessarily to have fun doing office work. Just like it's not necessarily the idea to have your emotions activate for video games instead of real life.

The idea is that once you construct an existence / life story that seems to make sense, then it's all right to bind emotions to the parts of that story, with strength proportional to their long-term impact. The anomie of today's world, where we simultaneously (a) engage in office work and (b) lack any passion in it, does not need to carry over: you should either fix one of those problems, or the other.

On a higher, more abstract level, this carries over the idiom of reinforcement over instrumental correlates of terminal values. In principle, this is something that a purer optimization process wouldn't do. You need neither happiness nor sadness to maximize expected utility. You only need to know which actions result in which consequences, and update that pure probability distribution as you learn through observation; something akin to "reinforcement" falls out of this, but without the risk of losing purposes, without any pleasure or pain. An agent like this is simpler than a human and more powerful - if you think that your emotions give you a supernatural advantage in optimization, you've entirely failed to understand the math of this domain. For a pure optimizer, the "advantage" of starting out with one more emotion bound to instrumental events is like being told one more abstract belief about which policies maximize expected utility, except that the belief is very hard to update based on further experience.

But it does not seem to me, that a mind which *has* the most value, is the same kind of mind that most *efficiently optimizes* values outside it. The interior of a true expected utility maximizer might be pretty boring, and I even suspect that you can build them to not be sentient.

For as far as my human eyes can see, I don't know what kind of mind I *should*value, if that mind lacks pleasure and happiness and emotion in the everyday events of its life. Bearing in mind that we are constructing this Future using our own preferences, not having it handed to us by some inscrutable external author.

If there's some better way of *being* (not just *doing*) that stands somewhere outside this, I have not yet understood it well enough to *prefer*it. But if so, then all this discussion of emotion would be as moot as it would be for an expected utility maximizer - one which was not valued at all for itself, but only valued for that which it maximized.

It's just hard to see why we would *want* to become something like that, bearing in mind that morality is not an inscrutable light handing down awful edicts from somewhere outside us.

At any rate - the hell of a life of disconnected episodes, where your actions don't connect strongly to anything you strongly care about, and nothing that you do all day invokes any passion - this angstseems avertible, however often it pops up in poorly written Utopias.

Serious Stories

****Every Utopia ever constructed - in philosophy, fiction, or religion - has been, to one degree or another, a place where you wouldn't *actually want* to live. I am not alone in this important observation: George Orwell said much the same thing in "Why Socialists Don't Believe In Fun", and I expect that many others said it earlier.

If you read books on How To Write - and there are a *lot* of books out there on How To Write, because amazingly a lot of book-writers think they know something about writing - these books will tell you that stories must contain "conflict".

That is, the more *lukewarm*sort of instructional book will tell you that stories contain "conflict". But some authors speak more plainly.

"Stories are about people's pain." Orson Scott Card.

"Every scene must end in disaster." Jack Bickham.

In the age of my youthful folly, I took for granted that *authors* were excused from the search for true Eutopia, because if you constructed a Utopia that *wasn't* flawed... what stories could you write, set there? "Once upon a time they lived happily ever after." What use would it be for a science-fiction author to try to depict a positive Singularity, when a positive Singularity would be...

...the end of all stories?

It seemed like a reasonable framework with which to examine the literary problem of Utopia, but something about that final conclusion produced a quiet, nagging doubt.

At that time I was thinking of an AI as being something like a safe wish-granting genie for the use of individuals. So the conclusion did make a kind of sense. If there was a problem, you would just wish it away, right? Ergo - no stories. So I ignored the quiet, nagging doubt.

Much later, after I concluded that even a safe genie wasn't such a good idea, it also seemed in retrospect that "no stories" could have been a productive indicator. On this particular occasion, "I can't think of a single story I'd *want to read* about this scenario", might indeed have pointed me toward the reason "I wouldn't want to *actually live* in this scenario".

So I swallowed my trained-in revulsion of Luddism and theodicy, and at least *tried*to contemplate the argument:

- A world in which nothing ever goes wrong, or no one ever experiences any pain or sorrow, is a world containing no stories worth reading about.
- A world that you wouldn't want to read about is a world where you wouldn't want to live.

- Into each eudaimonic life a little pain must fall. QED.

In one sense, it's clear that we do *not* want to live the sort of lives that are depicted in most stories that human authors have written so far. Think of the truly great stories, the ones that have become legendary for being the very best of the best of their genre: The *Iliiad, Romeo and Juliet, The Godfather, Watchmen,Planescape: Torment*, the second season of *Buffy the Vampire Slayer*, or*that ending*in *Tsukihime*. Is there a single story on the list that *isn't* tragic?

Ordinarily, we prefer pleasure to pain, joy to sadness, and life to death. Yet it seems we prefer to empathize with hurting, sad, dead characters. Or stories about happier people *aren't serious,*aren't artistically great enough to be worthy of praise - but then why selectively praise stories containing unhappy people? Is there some hidden benefit to us in it? It's a puzzle either way you look at it.

When I was a child I couldn't write fiction because I wrote things to go *well* for my characters - just like I wanted things to go well in real life. Which I was cured of by Orson Scott Card: *Oh,* I said to myself, *that's what I've been doing wrong, my characters aren't hurting.* Even then, I didn't realize that the microstructure of a plot works the same way - until Jack Bickham said that every scene must end in disaster. Here I'd been trying to set up problems and *resolve* them, instead of making them *worse...*

You simply don't *optimize*a story the way you optimize a real life. The *best*story and the *best*life will be produced by different criteria.

In the real world, people can go on living for quite a while without any major disasters, and still seem to do pretty okay. When was the last time you were shot at by assassins? Quite a while, right? Does your life seem emptier for it?

But on the other hand...

For some odd reason, when authors get too old or too successful, they revert to my childhood. Their stories start going *right.* They stop doing horrible things to their characters, with the result that they start doing horrible things to their readers. It seems to be a regular part of Elder Author Syndrome. Mercedes Lackey, Laurell K. Hamilton, Robert Heinlein, even Orson Scott bloody Card - they all went that way. They forgot how to hurt their characters. I don't know why.

And when you read a story by an Elder Author or a pure novice - a story where things just *relentlessly go right* one after another - where the main character defeats the supervillain with a snap of the fingers, or even worse, before the final battle, the supervillain *gives up and apologizes and then they're friends again* -

It's like a fingernail scraping on a blackboard at the base of your spine. If you've never actually read a story like that (or worse, written one) then count yourself lucky.

That fingernail-scraping quality - would it transfer over from the story to real life, if you tried living real life without a single drop of rain?

One answer might be that what a story really needs is not "disaster", or "pain", or even "conflict", but simply *striving.* That the problem with Mary Sue stories is that there's not enough striving in them, but they wouldn't actually need *pain.* This might, perhaps, be tested.

An alternative answer might be that this *is* the transhumanist version of Fun Theory we're talking about. So we can reply, "Modify brains to eliminate that fingernail-scraping feeling", unless there's some justification for keeping it. If the fingernail-scraping feeling is a pointless random bug getting in the way of Utopia, delete it.

Maybe we *should.* Maybe all the Great Stories are tragedies because... well...

I once read that in the BDSM community, "intense sensation" is a euphemism for pain. Upon reading this, it occurred to me that, the way humans are constructed now, it is just *easier* to produce pain than pleasure. Though I speak here somewhat outside my experience, I expect that it takes a highly talented and experienced sexual artist working for hours to produce a *good* feeling as intense as the pain of one strong kick in the testicles - which is doable in seconds by a novice.

Investigating the life of the priest and proto-rationalist Friedrich Spee von Langenfeld, who heard the confessions of accused witches, I looked up some of the instruments that had been used to produce confessions. There is no ordinary way to make a human being feel as *good* as those instruments would make you hurt. I'm not sure even drugs would do it, though my experience of drugs is as nonexistent as my experience of torture.

There's something imbalanced about that.

Yes, human beings are too optimistic in their planning. If losses weren't more aversive than gains, we'd go broke, the way we're constructed now. The experimental rule is that losing a desideratum - $50, a coffee mug, whatever - hurts between 2 and 2.5 times as much as the equivalent gain.

But this is a deeper imbalance than that. The effort-in/intensity-out difference between sex and torture is not a mere factor of 2.

If someone goes in search of sensation - in this world, the way human beings are constructed now - it's not surprising that they should arrive at pains to be mixed into their pleasures as a source of *intensity* in the combined experience.

If only people were constructed differently, so that you could produce pleasure as intense and in as many different flavors as pain! If only you could, with the same ingenuity and effort as a torturer of the Inquisition, make someone feel as *good* as the Inquisition's victims felt *bad* -

But then, what *is* the analogous pleasure that feels that good? A victim of skillful torture will do anything to stop the pain and anything to prevent it from being repeated. Is the equivalent pleasure one that overrides everything with

the demand to continue and repeat it? If people are stronger-willed to bear the pleasure, is it really the same pleasure?

There is another rule of writing which states that stories have to *shout*. A human brain is a long way off those printed letters. Every event and feeling needs to take place at ten times natural volume in order to have any impact at all. You must not try to make your characters behave or feel *realistically*- especially, you must not faithfully reproduce your own past experiences - because *without exaggeration*, they'll be too quiet to rise from the page.

Maybe all the Great Stories are tragedies because happiness can't shout loud enough - to a human reader.

Maybe that's what needs fixing.

And if it were fixed... would there be any use left for pain or sorrow? For even the *memory*of sadness, if all things were already as good as they could be, and every remediable ill already remedied?

Can you just delete pain outright? Or does removing the old floor of the utility function just create a new floor? Will any pleasure less than 10,000,000 hedons be the new unbearable pain?

Humans, built the way we are now, do seem to have hedonic scaling tendencies. Someone who can remember starving will appreciate a loaf of bread more than someone who's never known anything but cake. This was George Orwell's hypothesis for why Utopia is impossible in literature and reality:

"It would seem that human beings are not able to describe, nor perhaps to imagine, happiness except in terms of contrast... The inability of mankind to imagine happiness except in the form of relief, either from effort or pain, presents Socialists with a serious problem. Dickens can describe a poverty-stricken family tucking into a roast goose, and can make them appear happy; on the other hand, the inhabitants of perfect universes seem to have no spontaneous gaiety and are usually somewhat repulsive into the bargain."

For an expected utility maximizer, rescaling the utility function to add a trillion to all outcomes is meaningless - it's literally the same utility function, as a mathematical object. A utility function describes the *relative*intervals between outcomes; that's what it is, mathematically speaking.

But the human brain has distinct neural circuits for positive feedback and negative feedback, and different varieties of positive and negative feedback. There are people today who "suffer" from congenital analgesia - a total absence of pain. I never heard that *insufficient pleasure* becomes intolerable to them.

Congenital analgesics do have to inspect themselves carefully and frequently to see if they've cut themselves or burned a finger. Pain serves a purpose in the human mind design...

But that does not show there's no alternative which could serve the same purpose. Could you delete pain and replace it *with an urge not to do certain things*

that lacked the intolerable subjective quality of pain? I do not know all the Law that governs here, but I'd have to guess that yes, you could; you could replace that side of yourself with something more akin to an expected utility maximizer.

Could you delete the human tendency to scale pleasures - delete the accomodation, so that each new roast goose is as delightful as the last? I would guess that you could. This verges perilously close to deleting Boredom, which is right up there with Sympathy as an absolute indispensable... but to say that an old solution remains as pleasurable, is not to say that you will lose the urge to seek new and better solutions.

Can you make every roast goose as pleasurable as it would be in contrast to starvation, without ever having starved?

Can you prevent the pain of a dust speck irritating your eye from being the new torture, if you've literally *never experienced* anything *worse* than a dust speck irritating your eye?

Such questions begin to exceed my grasp of the Law, but I would guess that the answer is: yes, it can be done. It is my experience in such matters that once you do learn the Law, you can usually see how to do weird-seeming things.

So far as I know or can guess, David Pearce (*The Hedonistic Imperative*) is very probably right about the *feasibility* part, when he says:

"Nanotechnology and genetic engineering will abolish suffering in all sentient life. The abolitionist project is hugely ambitious but technically feasible. It is also instrumentally rational and morally urgent. The metabolic pathways of pain and malaise evolved because they served the fitness of our genes in the ancestral environment. They will be replaced by a different sort of neural architecture - a motivational system based on heritable gradients of bliss. States of sublime well-being are destined to become the genetically pre-programmed norm of mental health. It is predicted that the world's last unpleasant experience will be a precisely dateable event."

Is that... what we *want?*

To just wipe away the last tear, and be done?

Is there any good reason *not*to, except status quo bias and a handful of worn rationalizations?

What would be the *alternative?* Or alternatives?

To leave things as they are? Of course not. No God designed this world; we have no reason to think it exactly optimal on any dimension. If this world does not contain too much pain, then it must not contain enough, and the latter seems unlikely.

But perhaps...

You could cut out just the *intolerable*parts of pain?

Get rid of the Inquisition. Keep the sort of pain that tells you not to stick your finger in the fire, or the pain that tells you that you shouldn't have put your friend's finger in the fire, or even the pain of breaking up with a lover.

Try to get rid of the sort of pain that *grinds down and destroys* a mind. Or configure minds to be harder to damage.

You could have a world where there were broken legs, or even broken hearts, but no broken *people*. No child sexual abuse that turns out more abusers. No people ground down by weariness and drudging minor inconvenience to the point where they contemplate suicide. No random meaningless endless sorrows like starvation or AIDS.

And if even a broken leg still seems too scary -

Would we be less frightened of pain, if we were stronger, if our daily lives did not already exhaust so much of our reserves?

So that would be one alternative to the Pearce's world - if there are yet other alternatives, I haven't thought them through in any detail.

The path of courage, you might call it - the idea being that if you eliminate the destroying kind of pain and strengthen the people, then what's left shouldn't be *that* scary.

A world where there is sorrow, but not massive systematic *pointless*sorrow, like we see on the evening news. A world where pain, if it is not eliminated, at least does not *overbalance pleasure*. You could write stories about that world, and they could read our stories.

I do tend to be rather conservative around the notion of deleting large parts of human nature. I'm not sure how many major chunks you can delete until that balanced, conflicting, dynamic structure collapses into something simpler, like an expected pleasure maximizer.

And so I do admit that it is the path of courage that appeals to me.

Then again, I haven't lived it both ways.

Maybe I'm just *afraid* of a world so different as Analgesia - wouldn't that be an ironic reason to walk "the path of courage"?

Maybe the path of courage just seems like the *smaller change* - maybe I just have trouble empathizing over a larger gap.

But "change" is a moving target.

If a human child grew up in a *less* painful world - if they had never lived in a world of AIDS or cancer or slavery, and so did not know these things as evils that had *been triumphantly eliminated* - and so did not feel that they were "already done" or that the world was "already changed enough"...

Would they take the next step, and try to eliminate the unbearable pain of broken hearts, when someone's lover stops loving them?

And then what? Is there a point where *Romeo and Juliet* just seems less and less relevant, more and more a relic of some distant forgotten world? Does there come some point in the transhuman journey where the whole business of the negative reinforcement circuitry, can't possibly seem like anything except a pointless hangover to wake up from?

And if so, is there any point in *delaying*that last step? Or should we just throw away our fears and... throw away our fears?

I don't know.

Eutopia is Scary

"The big thing to remember about far-future cyberpunk is that it will be truly *ultra*-tech. The mind and body changes available to a 23rd-century Solid Citizen would probably amaze, disgust and *frighten* that 2050 netrunner!"
— *GURPS Cyberpunk*

Pick up someone from the 18th century - a *smart* someone. Ben Franklin, say. Drop them into the early 21st century.

We, in our time, think our life has improved in the last two or three hundred years. Ben Franklin is probably smart and forward-looking enough to *agree* that life has improved. But if you don't think Ben Franklin would be amazed, disgusted, and *frightened*, then I think you far overestimate the "normality" of your own time. You can think of reasons why Ben should find our world compatible, but Ben himself might not do the same.

Movies that were made in say the 40s or 50s, seem much more alien - to me - than modern movies allegedly set hundreds of years in the future, or in different universes. Watch a movie from 1950 and you may see a man slapping a woman. Doesn't happen a lot in *Lord of the Rings*, does it? Drop back to the 16th century and one popular entertainment was setting a cat on fire. Ever see *that* in any moving picture, no matter how "lowbrow"?

("But," you say, "that's showing how discomforting the Past's culture was, not how scary the Future is." Of which I wrote, "When we look over history, we see changes away from *absurd* conditions such as everyone being a peasant farmer and women not having the vote, toward *normal* conditions like a majority middle class and equal rights...")

Something about the Future will shock we 21st-century folk, if we were dropped in without slow adaptation. This is not because the Future is cold and gloomy - I am speaking of a positive, successful Future; the negative outcomes are probably just blank. Nor am I speaking of the idea that every Utopia has some dark hidden flaw. I am saying that the Future would discomfort us *because* it is better.

This is another piece of the puzzle for why no author seems to have ever succeeded in constructing a Utopia worth-a-damn. When they are out to depict how marvelous and wonderful the world could be, if only we would all be Marxists or Randians or let philosophers be kings... they try to depict the resulting outcome as *comforting* and *safe.*

Again, George Orwell from "Why Socialists Don't Believe In Fun":

> "In the last part, in contrast with disgusting Yahoos, we are shown the noble Houyhnhnms, intelligent horses who are free from human failings. Now these horses, for all their high character and unfailing common sense, are remarkably dreary creatures. Like the inhabitants of various other Utopias, they are chiefly concerned with avoiding fuss. They live uneventful, subdued, 'reasonable' lives, free not only from quarrels, disorder or insecurity of any kind, but also from 'passion', including physical love. They choose their mates on eugenic principles, avoid excesses of affection, and appear somewhat glad to die when their time comes."

One might consider, in particular contrast, Timothy Ferris's observation:

"What is the opposite of happiness? Sadness? No. Just as love and hate are two sides of the same coin, so are happiness and sadness. Crying out of happiness is a perfect illustration of this. The opposite of love is indifference, and the opposite of happiness is - here's the clincher - boredom...
The question you should be asking isn't 'What do I want?' or 'What are my goals?' but 'What would excite me?'
Remember - boredom is the enemy, not some abstract 'failure.'"

Utopia is reassuring, unsurprising, and dull.

Eutopia is scary.

I'm not talking here about evil means to a good end, I'm talking about the good outcomes *themselves.* That is the proper relation of the Future to the Past when things turn out *well*, as we would know very well from history if we'd actually lived it, rather than looking back with benefit of hindsight.

Now... I don't think you can actually *build the Future* on the basis of asking how to scare yourself. The vast majority of possible changes are in the direction of higher entropy; only a very few discomforts stem from things getting *better.*

"I shock you therefore I'm right" is one of the most *annoying* of all non-sequiturs, and we certainly don't want to go *there.*

But on a purely *literary* level... and bearing in mind that fiction is not reality, and fiction is not optimized the way we try to optimize reality...

I try to write fiction, now and then. More rarely, I finish a story. Even more rarely, I let someone else look at it.

Once I finally got to the point of thinking that maybe you *should* be able to write a story set in Eutopia, I tried doing it.

But I had something like an instinctive revulsion at the indulgence of trying to build a world that fit *me,* but probably wouldn't fit others so nicely.

So - without giving the world a seamy underside, or putting Knight Templars in charge, or anything so obvious as that - without deliberately trying to make the world*flawed*-

I was trying to invent, even if I had to do it myself, a better world where I would be *out of place.* Just like Ben Franklin would be out of place in the modern world.

Definitely *not* someplace that a transhumanist/science-advocate/libertarian (like myself) would go, and be smugly satisfied at how well all their ideas had worked. Down that path lay the Dark Side - certainly in a purely literary sense.

And you couldn't avert that just by having the Future go wrong in all the stupid obvious ways that transhumanists, or libertarians, or public advocates of science had already warned against. Then you just had a dystopia, and it might make a good SF story but it had already been done.

But I had my world's foundation, an absurd notion inspired by a corny pun; a vision of what you see when you wake up from cryonic suspension, that I couldn't have gotten away with posting to any transhumanist mailing list even as a joke.

And then, whenever I could think of an arguably-good idea that *offended my sensibilities*, I added it in. The goal being to - without ever deliberately making the Future *worse*- make it a place where I would be as shocked as possible to see that *that* was how things had turned out.

Getting rid of textbooks, for example - postulating that talking about science in public is socially unacceptable, for the same reason that you don't tell someone aiming to see a movie whether the hero dies at the end. A world that had rejected my beloved concept of science as the public knowledge of humankind.

Then I added up all the discomforting ideas together...

...and at least in my imagination, it worked better than anything I'd ever dared to visualize as a *serious*proposal.

My serious proposals had been optimized to look sober and safe and sane; everything voluntary, with clearly lighted exit signs, and all sorts of volume controls to prevent anything from getting too *loud* and waking up the neighbors. Nothing too absurd. Proposals that wouldn't scare the nervous, containing as little as possible that would cause anyone to make a fuss.

This world was ridiculous, and it was going to wake up the neighbors.

It was also seductive to the point that I had to exert a serious effort to prevent my soul from getting sucked out. (I suspect that's a general problem; that it's a good idea *emotionally*(not just *epistemically*) to *not*visualize your better Future in too much detail. You're better off comparing yourself to the Past. I may write a separate post on this.)

And so I found myself being pulled in the direction of this world in which I was supposed to be "out of place". I started thinking that, well, maybe it really *would* be a good idea to get rid of all the textbooks, all they do is take the fun out of science. I started thinking that maybe personal competition *was*a legitimate motivator (previously, I would have called it a zero-sum game and been morally aghast). I began to worry that peace, democracy, market economies, and con - but I'd better not finish that sentence. I started to wonder if the old vision that was so *reassuring,* so *safe,* was optimized to be good news to a modern human living in constant danger of permanent death or damage, and less optimized for the everyday existence of someone less frightened.

This is what happens when I try to invent a world that fails to confirm my sensibilities? It makes me wonder what would happen if someone else tried the same exercise.

Unfortunately, I can't seem to visualize any new world that represents the same shock to me as the last one did. Either the trick only works once, or you have to wait longer between attempts, or I'm too old now.

But I hope that so long as the world offends the *original* you, it gets to keep its literary integrity even if you start to find it less shocking.

I haven't yet published any story that gives more than a glimpse of this setting. I'm still debating with myself whether I dare. I don't know whether the suck-out-your-soul effect would threaten anyone but myself as author - I haven't seen it happening with Banks's Culture or Wright's Golden Oecumene, so I suspect it's more of a trap when a world fits a single person too well. But I got enough flak when I presented the case for getting rid of textbooks.

Still - I have seen the possibilities, now. So long as no one dies permanently, I am leaning in favor of a loud and scary Future.**

Building Weirdtopia

"Two roads diverged in the woods. I took the one less traveled, and had to eat bugs until Park rangers rescued me."
— Jim Rosenberg

Utopia and Dystopia have something in common: they both confirm the moral sensibilities you started with. Whether the world is a libertarian utopia of the non-initiation of violence and everyone free to start their own business, or a hellish dystopia of government regulation and intrusion - you might like to find

yourself in the first, and hate to find yourself in the second; but either way you nod and say, "Guess I was right all along."

So as an exercise in creativity, try writing them down side by side: Utopia, Dystopia, and Weirdtopia. The zig, the zag and the zog.

I'll start off with a worked example for *public understanding of science:*

- *Utopia:* Most people have the equivalent of an undergrad degree in something; *everyone* reads the popular science books (and they're *good*books); everyone over the age of nine understands evolutionary theory and Newtonian physics; scientists who make major contributions are publicly adulated like rock stars.
- *Dystopia:* Science is considered boring and possibly treasonous; public discourse elevates religion or crackpot theories; stem cell research is banned.
- *Weirdtopia:* Science is kept secret to avoid spoiling the surprises; no public discussion but intense private pursuit; cooperative ventures surrounded by fearsome initiation rituals because that's what it takes for people to feel like they've actually learned a Secret of the Universe and be satisfied; someone you meet may only know extremely basic science, but they'll have personally done revolutionary-level work in it, just like you. Too bad you can't compare notes.

Disclaimer 1: Not every sensibility we have is necessarily *wrong.* Originality is a goal of literature, not science; sometimes it's better to be right than to be new. But there are also such things as cached thoughts. At least in my own case, it turned out that trying to invent a world that went outside my pre-existing sensibilities, did me a world of good.

Disclaimer 2: This method is not universal: Not all interesting ideas fit this mold, and not all ideas that fit this mold are good ones. Still, it seems like an interesting technique.

If you're trying to write science fiction (where originality *is* a legitimate goal), then you can write down anything nonobvious for Weirdtopia, and you're done.

If you're trying to do Fun Theory, you have to come up with a Weirdtopia that's at least *arguably-better*than Utopia. This is harder but also directs you to more interesting regions of the answer space.

If you can make all your answers *coherent* with each other, you'll have quite a story setting on your hands. (Hope you know how to handle characterization, dialogue, description, conflict, and all that other stuff.)

Here's some partially completed challenges, where I wrote down a Utopia and a Dystopia (according to the moral sensibilities I started with before *I* did this exercise), but inventing a (better) Weirdtopia is left to the reader.

Economic...

- *Utopia:* The world is flat and ultra-efficient. Prices fall as standards of living rise, thanks to economies of scale. Anyone can easily start their own business and most people do. Everything is done in the right place by the right person under Ricardo's Law of Comparative Advantage. Shocks are efficiently absorbed by the risk capital that insured them.
- *Dystopia:* Lots of trade barriers and subsidies; corporations exploit the regulatory systems to create new barriers to entry; dysfunctional financial systems with poor incentives and lots of unproductive investments; rampant agent failures and systemic vulnerabilities; standards of living flat or dropping.
- *Weirdtopia:* _____

Sexual...

- *Utopia:* Sexual mores straight out of a Spider Robinson novel: Sexual jealousy has been eliminated; no one is embarrassed about what turns them on; universal tolerance and respect; everyone is bisexual, poly, and a switch; total equality between the sexes; no one would look askance on sex in public any more than eating in public, so long as the participants cleaned up after themselves.
- *Dystopia:* 10% of women have never had an orgasm. States adopt laws to ban gay marriage. Prostitution illegal.
- *Weirdtopia:* _____

Governmental...

- *Utopia:* Non-initiation of violence is the chief rule. Remaining public issues are settled by democracy: Well reasoned public debate in which all sides get a free voice, followed by direct or representative majority vote. Smoothly interfunctioning Privately Produced Law, which coordinate to enforce a very few global rules like "no slavery".
- *Dystopia:* Tyranny of a single individual or oligarchy. Politicians with effective locks on power thanks to corrupted electronic voting systems, voter intimidation, voting systems designed to create coordination problems. Business of government is unpleasant and not very competitive; hard to move from one region to another.
- *Weirdtopia:* _____

Technological...

- *Utopia:* All Kurzweilian prophecies come true simultaneously. Every pot contains a chicken, a nanomedical package, a personal spaceship, a superdupercomputer, amazing video games, and a pet AI to help you use it all, plus a pony. Everything is designed by Apple.
- *Dystopia:* Those damned fools in the government banned everything more complicated than a lawnmower, and we couldn't use our lawnmowers after Peak Oil hit.
- *Weirdtopia:* _____

Cognitive...

- *Utopia:* Brain-computer implants for everyone! You can do whatever you like with them, it's all voluntary and the dangerous buttons are clearly labeled. There are AIs around that are way more powerful than you; but they don't hurt you unless you ask to be hurt, sign an informed consent release form and click "Yes" three times.
- *Dystopia:* The first self-improving AI was poorly designed, everyone's dead and the universe is being turned into paperclips. Or the augmented humans hate the normals. Or augmentations make you go nuts. Or the darned government banned everything again, and people are still getting Alzheimers due to lack of stem-cell research.
- *Weirdtopia:* _____

Justified Expectation of Pleasant Surprises

I recently tried playing a computer game that made a major fun-theoretic error. (At least I strongly suspect it's an error, though they are game designers and I am not.)

The game showed me - right from the start of play - what abilities I could purchase as I increased in level. Worse, there were many different choices; still worse, you had to pay a cost in fungible points to acquire them, making you feel like you were losing a resource... But today, I'd just like to focus on the problem of telling me, *right at the start of the game,* about all the nice things that might happen to me later.

I can't think of a good experimental result that backs this up; but I'd expect that a pleasant *surprise* would have a greater hedonic impact, than being told about the same gift in advance. Sure, the moment you were first *told*about the gift would be *good news*, a moment of pleasure in the moment of being

told. But you wouldn't have the gift in hand at that moment, which limits the pleasure. And then you have to wait. And then when you finally get the gift - it's pleasant to go from not having it to having it, *if* you didn't wait too long; but a surprise would have a larger momentary impact, I would think.

This particular game had a status screen that showed *all* my future class abilities *at the start of the game* - inactive and dark but with full information still displayed. From a hedonic standpoint this seems like *miserable*fun theory. All the "good news" is lumped into a gigantic package; the items of news would have much greater impact if encountered separately. And then I have to wait a long time to actually acquire the abilities, so I get an extended period of comparing my current weak game-self to all the wonderful abilities I *could* have but don't.

Imagine living in two possible worlds. Both worlds are otherwise rich in challenge, novelty, and other aspects of Fun. In both worlds, you get smarter with age and acquire more abilities over time, so that your life is always getting better.

But in *one*world, the abilities that come with seniority are openly discussed, hence widely known; you know what you have to look forward to.

In the *other*world, anyone older than you will *refuse to talk* about certain aspects of growing up; you'll just have to wait and find out.

I ask you to contemplate - not just which world you might prefer to live in - but how *much* you might want to live in the second world, rather than the first. I would even say that the second world seems more *alive;* when I imagine living there, my imagined *will to live* feels stronger. I've got to stay alive to find out what happens next, right?

The idea that *hope* is important to a happy life, is hardly original with me - though I think it might not be emphasized quite *enough*, on the lists of things people are told they need.

I don't agree with buying lottery tickets, but I do think I understand why people do it. I remember the times in my life when I had more or less belief that things would improve - that they were heading up in the near-term or mid-term, close enough to anticipate. I'm having trouble describing how much of a difference it makes. Maybe I don't *need* to describe that difference, unless some of my readers have never had any light at the end of their tunnels, or some of my readers have never looked forward and seen darkness.

If existential angst comes from having at least one deep problem in your life that you aren't thinking about explicitly, so that the pain which comes from it seems like a natural permanent feature - then the very first question I'd ask, to identify a possible source of that problem, would be, "Do you expect your life to improve in the near or mid-term future?"

Sometimes I meet people who've been run over by life, in much the same way as being run over by a truck. Grand catastrophe isn't necessary to destroy a will to live. The extended absence of hope leaves the same sort of wreckage.

People need hope. I'm not the first to say it.

But I think that the importance of *vague hope* is underemphasized.

"Vague" is usually not a compliment among rationalists. Hear "vague hopes" and you immediately think of, say, an alternative medicine herbal profusion whose touted benefits are so conveniently unobservable (not to mention experimentally unverified) that people will buy it for anything and then refuse to admit it didn't work. You think of poorly worked-out plans with missing steps, or supernatural prophecies made carefully unfalsifiable, or fantasies of unearned riches, or...

But you know, generally speaking, our beliefs about the future *should* be vaguer than our beliefs about the past. We just know less about tomorrow than we do about yesterday.

There are plenty of *bad* reasons to be vague, all sorts of *suspicious*reasons to offer nonspecific predictions, but reversed stupidity is not intelligence: When you've eliminated all the ulterior motives for vagueness, your beliefs about the future should *still*be vague.

We don't know much about the future; let's hope *that*doesn't change for as long as human emotions stay what they are. Of all the poisoned gifts a big mind could give a small one, a walkthrough for the game has to be near the top of the list.

What we need to maintain our interest in life, is a *justified expectation of pleasant surprises.* (And yes, you can expect a surprise if you're not logically omniscient.) This excludes the herbal profusions, the poorly worked-out plans, and the supernatural. The best reason for this justified expectation is *experience*, that is, being pleasantly surprised on a frequent yet irregular basis. (If this isn't happening to you, please file a bug report with the appropriate authorities.)

Vague justifications for believing in a pleasant *specific outcome* would be the opposite.

There's also other dangers of having pleasant hopes that are *too specific* - even if *justified*, though more often they aren't - and I plan to talk about that in the next post.

Seduced by Imagination

"Vagueness" usually has a bad name in rationality - connoting skipped steps in reasoning and attempts to avoid falsification. But a rational view of the Future *should*be vague, because the information we have about the Future is weak. Yesterday I argued that justified vague hopes might also be better *hedonically* than specific foreknowledge - the power of pleasant surprises.

But there's also a more severe warning that I must deliver: It's not a good idea to dwell much *on* imagined pleasant futures, since you can't actually dwell *in*them. It can suck the emotional energy out of your actual, current, ongoing life.

Epistemically, we know the Past much more *specifically*than the Future. But also on *emotional* grounds, it's probably wiser to compare yourself to Earth's past, so you can see how far we've come, and how much better we're doing. Rather than comparing your life to an imagined future, and thinking about how awful you've got it Now.

Having set out to explain George Orwell's observation that no one can seem to write about a Utopia where anyone would want to live - having laid out the various Laws of Fun that I believe are being *violated*in these dreary Heavens - I am now explaining why you *shouldn't* apply this knowledge to invent an extremely seductive Utopia and write stories set there. That may suck out your soul like an emotional vacuum cleaner.

I briefly remarked on this phenomenon earlier, and someone said, "Define 'suck out your soul'." Well, it's mainly a tactile thing: you can practically *feel*the pulling sensation, if your dreams wander too far into the Future. It's like something out of H. P. Lovecraft: *The Call of Eutopia.* A professional hazard of having to stare out into vistas that *humans were meant to gaze upon*, and knowing *a little too much* about the lighter side of existence.

But for the record, I will now lay out the components of "soul-sucking", that you may recognize the bright abyss and steer your thoughts away:

- Your emotional energy drains away into your imagination of Paradise:
  - You find yourself thinking of it more and more often.
  - The actual challenges of your current existence start to seem less interesting, less compelling; you think of them less and less.
  - Comparing everything to your imagined perfect world heightens your annoyances and diminishes your pleasures.
- You go into an affective death spiral around your imagined scenario; you're reluctant to admit anything bad could happen on your assumptions, and you find more and more nice things to say.
- Your mind begins to forget the difference between fiction and real life:
  - You originally made many arbitrary or iffy choices in constructing your scenario. You forget that the Future is actually more unpredictable than this, and that you made your choices using limited foresight and merely human optimizing ability.
  - You forget that, in real life, at least *some*of your amazing good ideas are *guaranteed*not to work as well as they do in your imagination.

– You start wanting the *exact specific* Paradise you imagined, and worrying about the disappointment if you don't get that *exact* thing.

Hope can be a dangerous thing. And when you've just been hit hard - at the moment when you most *need* hope to keep you going - that's also when the real world seems most painful, and the world of imagination becomes most seductive.

It's a balancing act, I think. One needs enough Fun Theory to truly and legitimately justify hope in the future. But not a detailed vision so seductive that it steals emotional energy from the real life and real challenge of creating that future. You need "a light at the end of the secular rationalist tunnel" as Roko put it, but you don't want people to drift away from their bodies into that light.

So *how much* light is that, exactly? Ah, now that's the issue.

I'll start with a simple and genuine question: Is what I've already said, enough?

Is knowing the abstract fun theory and being able to pinpoint the exact flaws in previous flawed Utopias, enough to make you look forward to tomorrow? Is it enough to inspire a stronger will to live? To dispel worries about a long dark tea-time of the soul? Does it now seem - on a gut level - that if we could really build an AI and really shape it, the resulting future would be very much worth staying alive to see?

The Uses of Fun (Theory)

"But is there anyone who actually wants to live in a Wellsian Utopia? On the contrary, not to live in a world like that, not to wake up in a hygenic garden suburb infested by naked schoolmarms, has actually become a conscious political motive. A book like Brave New World is an expression of the actual fear that modern man feels of the rationalised hedonistic society which it is within his power to create."
— George Orwell, *Why Socialists Don't Believe in Fun*

There are three reasons I'm talking about Fun Theory, some more important than others:

1. If every picture ever drawn of the Future looks like a terrible place to actually live, it might tend to drain off the motivation to create the future. It takes hope to sign up for cryonics.
2. People who leave their religions, but don't familiarize themselves with the deep, foundational, fully general arguments against theism, are at risk of backsliding. Fun Theory lets you look at our present world, and see that it is not optimized *even for considerations like personal responsibility or self-reliance.* It is the fully general reply to theodicy.

3. Going into the details of Fun Theory helps you see that eudaimonia is actually *complicated*- that there are a lot of properties necessary for a mind to lead a worthwhile existence. Which helps you appreciate just how worthless a galaxy would end up looking (with extremely high probability) if it was optimized by something with a utility function rolled up at random.

To amplify on these points in order:

(1) You've got folks like Leon Kass and the other members of Bush's "President's Council on Bioethics" running around talking about what a terrible, terrible thing it would be if people lived longer than threescore and ten. While some philosophers have pointed out the flaws in their arguments, it's one thing to point out a flaw and another to provide a counterexample. "Millions long for immortality who do not know what to do with themselves on a rainy Sunday afternoon," said Susan Ertz, and that argument will sound plausible for as long as you can't imagine what to do on a rainy Sunday afternoon, and it seems unlikely that anyone could imagine it.

It's not exactly the fault of Hans Moravec that his world in which humans are kept by superintelligences as pets, doesn't sound quite Utopian. Utopias are just really hard to construct, for reasons I'll talk about in more detail later - but this observation has already been made by many, including George Orwell.

Building the Future is part of the ethos of secular humanism, our common project. If you have nothing to look forward to - if there's no image of the Future that can inspire real enthusiasm - then you won't be able to scrape up enthusiasm for that common project. And if the project is, in fact, a worthwhile one, the expected utility of the future will suffer accordingly from that nonparticipation. So that's one side of the coin, just as the other side is living so exclusively in a fantasy of the Future that you can't bring yourself to go on in the Present.

I recommend thinking vaguely of the Future's hopes, thinking specifically of the Past's horrors, and spending *most*of your time in the Present. This strategy has certain epistemic virtues beyond its use in cheering yourself up.

But it helps to have *legitimate*reason to vaguely hope - to minimize the leaps of abstract optimism involved in thinking that, yes, you can live and obtain happiness in the Future.

(2) Rationality is our goal, and atheism is just a side effect - the judgment that happens to be produced. But atheism is an *important* side effect. John C. Wright, who wrote the heavily transhumanist *The Golden Age,* had some kind of temporal lobe epileptic fit and became a Christian. There's a once-helpful soul, now lost to us.

But it is possible to do better, even if your brain malfunctions on you. I know a transhumanist who has strong religious visions, which she once attributed

to future minds reaching back in time and talking to her... but then she reasoned it out, asking why future superminds would grant *only her* the solace of conversation, and why they could offer vaguely reassuring arguments but not tell her winning lottery numbers or the 900th digit of pi. So now she still has strong religious experiences, but she is not religious. That's the difference between weak rationality and strong rationality, and it has to do with the *depth* and *generality* of the epistemic rules that you know and apply.

Fun Theory is part of the fully general reply to religion; in particular, it is the fully general reply to theodicy. If you can't say how God could have *better* created the world without sliding into an antiseptic Wellsian Utopia, you can't carry Epicurus's argument. If, on the other hand, you have some idea of how you could build a world that was not only more pleasant but also a better medium for self-reliance, then you can see that permanently losing both your legs in a car accident when someone else crashes into you, doesn't seem very eudaimonic.

If we can imagine what the world might look like if it had been designed by anything remotely like a benevolently inclined superagent, we can look at the world around us, and see that *this*isn't it. This doesn't require that we correctly forecast the *full* optimization of a superagent - just that we can envision *strict improvements* on the present world, even if they prove not to be *maximal*.

(3) There's a severe problem in which people, due to anthropomorphic optimism and the lack of specific reflective knowledge about their invisible background framework and many other biases which I have discussed, think of a "nonhuman future" and just subtract off a few aspects of humanity that are salient, like enjoying the taste of peanut butter or something. While still envisioning a future filled with minds that have aesthetic sensibilities, experience happiness on fulfilling a task, get bored with doing the same thing repeatedly, etcetera. These things seem *universal,* rather than *specifically human* - to a human, that is. They don't involve having ten fingers or two eyes, so they must be universal, right?

And if you're still in this frame of mind - where "real values" are the ones that persuade every possible mind, and the rest is just some extra specifically human stuff - then Friendly AI will seem unnecessary to you, because, in its absence, you expect the universe to be *valuable* but not *human*.

It turns out, though, that once you start talking about what specifically is and isn't *valuable*, even if you try to keep yourself sounding as "non-human" as possible - then you still end up with a big complicated computation that is only instantiated physically in human brains and nowhere else in the universe. Complex challenges? Novelty? Individualism? Self-awareness? Experienced happiness? A paperclip maximizer cares not about these things.

It is a long project to crack people's brains loose of thinking that things will turn out regardless - that they can subtract off a few specifically human-seeming things, and then end up with plenty of other things they care about that are

universal and will appeal to arbitrarily constructed AIs. And of this I have said a very great deal already. But it does not seem to be enough. So Fun Theory is one more step - taking the curtains off some of the invisible background of our values, and revealing some of the complex criteria that go into a life worth living.

Higher Purpose

Long-time readers will recall that I've long been uncomfortable with the idea that you can adopt a Cause as a hedonic accessory:

"Unhappy people are told that they need a 'purpose in life', so they should pick out an altruistic cause that goes well with their personality, like picking out nice living-room drapes, and this will brighten up their days by adding some color, like nice living-room drapes."

But conversely it's also a fact that having a Purpose In Life consistently shows up as something that increases happiness, as measured by reported subjective well-being.

One presumes that this works equally well *hedonically*no matter how *misguided*that Purpose In Life may be - no matter if it is actually doing harm - no matter if the means are as cheap as prayer. Presumably, all that matters for *your*happiness is that *you*believe in it. So you had better not question overmuch whether you're really being effective; that would disturb the warm glow of satisfaction you paid for.

And here we verge on Zen, because you can't deliberately pursue "a purpose that takes you outside yourself", *in order to take yourself outside yourself.* That's still all about *you.*

Which is the whole Western concept of "spirituality" that I despise: *You* need a higher purpose so that *you* can be emotionally healthy. The external world is just a stream of victims for *you*to rescue.

Which is not to say that you can become more noble by being less happy. To *deliberately*sacrifice more, so that *you*can appear more virtuous to yourself, is also not a purpose outside yourself.

The way someone ends up with a real purpose outside themselves, is that they're walking along one day and see an elderly women passing by, and they realize "Oh crap, a hundred thousand people are dying of old age every day, what an awful way to die" and then they set out to do something about it.

If you want a purpose like that, then by *wanting* it, you're just circling back into *yourself*again. Thinking about *your*need to be "useful". Stop searching for *your purpose.* Turn your eyes *outward*to look at things outside yourself, and *notice*when you care about them; and then figure out how to be *effective,*instead of priding yourself on how much spiritual benefit you're getting just for trying.

With that said:

In today's world, most of the highest-priority legitimate Causes are about large groups of people in extreme jeopardy. (Wide scope * high severity.) Aging threatens the old, starvation threatens the poor, existential risks threaten humankind as a whole.

But some of the potential *solutions*on the table are, arguably, so powerful that they could solve almost the entire list. Some argue that nanotechnology would take almost all our current problems off the table. (I agree but reply that nanotech would create other problems, like unstable military balances, crazy uploads, and brute-forced AI.)

I sometimes describe the *purpose*(if not the actual decision criterion) of Friendly superintelligence as "Fix all fixable problems such that it is more important for the problem to be fixed *immediately*than fixed by our own efforts."

Wouldn't you then run out of victims with which to feed your higher purpose?

"Good," you say, "I should sooner step in front of a train, than ask that there be more victims just to keep myself occupied."

But do altruists then have little to look forward to, in the Future? Will we, deprived of victims, find our higher purpose shriveling away, and have to make a new life for ourselves as self-absorbed creatures?

"That unhappiness is relatively small compared to the unhappiness of a mother watching their child die, so screw it."

Well, but like it or not, the presence or absence of higher purpose *does* have hedonic effects on human beings, configured as we are now. And to reconfigure ourselves so that we no longer need to care about anything outside ourselves... does sound a little sad. I don't practice altruism *for the sake of being virtuous* - but I also recognize that "altruism" is part of what I value in humanity, part of what I want to save. If you save everyone, have you obsoleted the will to save them?

But I think it's a false dilemma. Right now, in this world, any halfway capable rationalist who looks outside themselves, will find their eyes immediately drawn to large groups of people in extreme jeopardy. Wide scope * great severity = big problem. It doesn't mean that if one were to *solve*all those Big Problems, we would have nothing *left*to care about except ourselves.

Friends? Family? Sure, and also more abstract ideals, like Truth or Art or Freedom. The change that altruists may have to get used to, is the absence of any solvable problems so *urgent* that it doesn't matter whether they're solved by a person or an unperson. That *is*a change and a major one - which I am not going to go into, because we don't yet live in that world. But it's not so sad a change, as having nothing to care about outside yourself. It's not the end of purpose. It's not even a descent into "spirituality": people might

still look around outside themselves to see what needs doing, thinking more of effectiveness than of emotional benefits.

But I *will*say this much:

If all goes well, there will come a time when you could search the whole of civilization and never find a single person so much in need of help, as dozens you now pass on the street.

If you do want to save someone from death, or help a great many people - if you want to have that memory for yourself, later - then you'd better get your licks in *now.*

I say this, because although that is not the *purest* motive, it is a *useful* motivation.

And for now - in this world - it is the *usefulness*that matters. That is the Art we are to practice *today*, however we imagine the world might change tomorrow. We are not here to be our hypothetical selves of tomorrow. *This*world is our charge and our challenge; we must adapt ourselves to live *here*, not somewhere else.

After all - to care whether your motives were sufficiently "high", would just turn your eyes inward.