MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Coherent Extrapolated Volition

Eliezer Yudkowsky
Machine Intelligence Research Institute

Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition*. The Singularity Institute, San Francisco, CA.

The Machine Intelligence Research Institute was previously known as the Singularity Institute.

1. Introduction

This is an update to that part of Friendly AI theory that describes Friendliness, the objective or thing-we're-trying-to-do. The information is current as of May 2004, and should not become dreadfully obsolete until late June, when I plan to have an unexpected insight. (Update: Actually, it took two days. Still, the text here isn't too far from the mark.)

Misleading terminology alert: I am still calling the Friendly Thingy an "Artificial Intelligence" or "superintelligence," even though it would be more accurate to call it a Friendly Really Powerful Optimization Process.

Warning: Beware of things that are fun to argue.

Defining Friendliness is not the life-or-death problem on which the survival of humanity depends. It is *a* life-or-death problem, but not *the* life-or-death problem. Friendly AI requires:

1. Solving the technical problems required to maintain a well-specified abstract invariant in a self-modifying goal system. (Interestingly, this problem is relatively straightforward from a theoretical standpoint.)
2. Choosing something nice to do with the AI. This is about midway in theoretical hairiness between problems 1 and 3.
3. Designing a framework for an abstract invariant that doesn't *automatically* wipe out the human species. *This is the hard part.*

Some examples to give a taste of the hard part:

- An AI whose abstract invariant is a "reward button," pressed when the AI pleases users, may go on to tile the solar system with ever-larger floating-point representations of total reward.
- An AI trained on the problem of "making humans smile" may acquire a utility function satisfied by tiling the solar system with tiny smiley faces.

Compare Friendly AI to the easier problem of wiping out the human species. The latter requires:

- Probing around in AI design space to find a few things that work (not necessarily understanding *why* they work) and tacking them together to do recursive self-improvement.
- Designing a coherent system based on self-modification which is probabilistic in nature and does not maintain an invariant optimization target. For example, directed evolution.

- Using a theory of intelligence which allows for *deductive* (probability ∼100%) recursive self-improvement that maintains an abstract invariant, *without* achieving the technical understanding to choose an invariant that does what the programmer thinks it does.
- Brute-forcing the problem. Moore's Law continually advances, increasing the computing power available to blind probers, decreasing the technical brilliance required to wipe out the human species. (More computing power helps not at all on FAI.)

Arguing about Friendliness is easy, fun, and distracting. Without a technical solution to FAI, it doesn't *matter* what the would-be designer of a superintelligence wants; those intentions will be irrelevant to the outcome. Arguing over Friendliness *content* is planning the Victory Party After The Revolution—not just before winning the battle for Friendly AI, but before there is any prospect of the human species putting up a fight before we go down. The goal is not to put up a good fight, but to *win*, which is much harder. But *right now* the question is whether the human species can field a non-pathetic force in defense of six billion lives and futures.

Suppose we get the funding, find the people, pull together the project, solve the technical problems of AI and Friendly AI, carry out the required teaching, and finally set a superintelligence in motion, all before anyone else throws something together that only does recursive self-improvement. It is still pointless if we do not have a nice task for that optimization process to carry out. You can fail just as easily on the last step of the problem as on the first. Failing on the last step of the problem is rarer, not because the last step of the problem is less difficult, but because to fail on the last step of the problem you must first succeed on all the other steps. Plenty of would-be revolutionaries spend all their days in cafes planning the New World Order, arguing moral philosophy, and writing constitutions for their lovely new government. But of those revolutions that *do* succeed, most fail on that last part of the problem—being nice—and go on to wreak havoc. To win you need to win on *every* critical stage of the problem.

Friendliness is the end; FAI theory is the means. Friendliness is the easiest part of the problem to explain—the part that says what we *want*. Like explaining why you want to fly to London, versus explaining a Boeing 747; explaining toast, versus explaining a toaster oven. Friendliness isn't the hardest part of the problem, or the one we need to solve *right now*, but all attention tends to focus on that which is easiest to argue about.

I'm writing this essay because I must—because it's one of my responsibilities to describe my current plans, should I succeed in solving the more difficult parts of the problem. Just because something is necessary doesn't make it less dangerous; just because my ethics require me to do a thing does not absolve me of the consequence. Beware lest Friendliness eat your soul.

2. Introducing Volition

Suppose you're faced with a choice between two boxes, A and B. One and only one of the boxes contains a diamond. You guess that the box which contains the diamond is box A. It turns out that the diamond is in box B. Your *decision* will be to take box A. I now apply the term *volition* to describe the sense in which you may be said to *want* box B, even though your guess leads you to pick box A.

Let's say that Fred wants a diamond, and Fred asks me to give him box A. I know that Fred wants a diamond, and I know that the diamond is in box B, and I want to be helpful. I could advise Fred to ask for box B instead; open up the boxes and let Fred look inside; hand box B to Fred; destroy box A with a flamethrower; quietly take the diamond out of box B and put it into box A; or let Fred make his own mistakes, to teach Fred care in choosing future boxes.

But I do not simply say: "Well, Fred chose box A, and he got box A, so I fail to see why there is a problem." There are several ways of stating my perceived problem:

1. Fred was disappointed on opening box A, and would have been happier on opening box B.
2. It is possible to predict that if Fred chooses box A, Fred will look back and wish he had chosen box B instead; while if Fred chooses box B, Fred will be satisfied with his choice.
3. Fred wanted "the box containing the diamond," not "box A," and chose box A only because he guessed that box A contained the diamond.
4. If Fred had known the correct answer to the question of simple fact, "Which box contains the diamond?", Fred would have chosen box B.

Hence my intuitive sense that giving Fred box A, as he literally requested, is not actually *helping* Fred.

If you find a genie bottle that gives you three wishes, it's probably a good idea to seal the genie bottle in a locked safety box under your bed, unless the genie pays attention to your volition, not just your decision.

Let's ask Fred's volition a more complicated question: "What should the code of a Friendly AI look like?" But there's a minor problem, which is that Fred is a hunter-gatherer from the Islets of Langerhans, where they have never even heard of Artificial Intelligence. There is a tremendous distance between the real Fred, and a Fred who might specify the code of a Friendly AI. Fred needs more than knowledge of simple facts, and knowledge of many background disciplines not known to the Islets of Langerhans. Fred needs to answer moral questions he doesn't know how to ask, acquire new skills, ponder possible future courses of the human species, and perhaps become smarter. If I started with the Fred of this passing moment, and began a crash education course which Fred absorbed successfully, it would be years before Fred could write a Friendly AI. Different Everett branches of Fred would not write identical code, and would probably write different AIs. Would today's Fred recognize any of those future Freds, who had learned so much about the universe, and rejected so many prevailing beliefs of the Islets of Langerhans?

Yet today's Fred still has wishes that straightforwardly apply to the creation of an Artificial Intelligence. Fred may not know what a paperclip is, or what a solar system is, but nonetheless Fred would not want anyone to create an Artificial Intelligence that tiled the solar system with paperclips.

Fred choosing between box A and box B is a simple case of asking what Fred "would want." The Fred who knows that box B contains the diamond is so close to today's Fred, so readily understandable, that we skip over the intervening step of extrapolation, and say simply: "Fred wants box B." Rather than saying, "I have mentally extrapolated a volition for Fred, consisting of an alternate version of Fred that knows which box contains the diamond, and I imagine that this Fred chooses box B."

Sadly, there isn't a little display built into the back of your neck that shows your volition. To construe your volition, I need to define a dynamic for extrapolating your volition, given knowledge about you. In the case of an FAI, this knowledge might include a complete readout of your brain-state, or an approximate model of your mind-state. The FAI takes the knowledge of Fred's brainstate, and other knowledge possessed by the FAI (such as which box contains the diamond), does . . . something complicated . . . and out pops a construal of Fred's volition.

I shall refer to the "something complicated" as the *dynamic*.

Note that *labeling* something "a volition-extrapolating dynamic" does not mean that it does anything of the sort. I can apply the term "volition-extrapolating dynamic" to a process that visualizes Fred replaced with a potted plant and decodes the waving leaves. But I think Fred would object, and I'm not just saying that because I imagine his leaves waving westward.

In all cases we speak of choosing and teaching a *generally intelligent AI* to extrapolate volition. If one attempted to write an ordinary computer program using ordinary computer programming skills, the task would be a thousand lightyears beyond hopeless.

2.1. Spread, Muddle, and Distance

These quantities are ad hoc and will probably be replaced by more elegant first principles when I have worked out more of the theory.

Spread describes cases where your extrapolated volition becomes unpredictable, intractable, or random. You might predictably want a banana tomorrow, or predictably not want a banana tomorrow, or predictably have a 30% chance of wanting a banana tomorrow depending on variables that are quantum-random, deterministic but unknown, or computationally intractable. When multiple outcomes are possible and probable, this creates *spread* in your extrapolated volition.

Muddle measures self-contradiction, inconsistency, and cases of "damned if you do and damned if you don't." Suppose that if you got a banana tomorrow you would not want a banana, and if you didn't get a banana you would indignantly complain that you wanted a banana. This is *muddle*.

Distance measures how difficult it would be to explain your volition to your current self, and the degree to which the volition was extrapolated by firm steps.

Short distance: An extrapolated volition that you would readily agree with if explained.

Medium distance: An extrapolated volition that would require extended education and argument before it became massively obvious in retrospect.

Long distance: An extrapolated volition your present-day self finds *incomprehensible*; not outrageous or annoying, but blankly incomprehensible.

Ground zero: Your actual decision.

2.2. Obvious Moral Hazard of Volitionism as Philosophy

In everyday life, a helpful human should not try to extrapolate a friend's volition beyond the point where the friend would readily agree. Our ability to extrapolate is too unreliable, and our evolved tendency to mess with other people's lives "for their own good" is too strong. If Fred says "I've heard the argument and I still disagree," that settles it; I don't think one should believe a politician who claims to follow Fred's volition when Fred himself protests. Humans dealing with other humans can usually essay only *short-distance* extrapolations and still claim to work on the other's behalf. Society does recognize dangerous exceptions to this rule, such as parents and children.

But I don't think it would be "helping" Fred to gift him with a genie which would try to explain to Fred why his wish was dangerous, but would go ahead and fulfill the wish, after Fred of the Islets indignantly instructed the genie to stop giving him insolent backtalk and follow orders. In this case I don't see much of a moral distinction between murder by genie bottle and murder by bullet.

Genies ain't human, humans ain't genies. Different precautions apply.

3. Coherent Extrapolated Volition

As of May 2004, my take on Friendliness is that the initial dynamic should implement the *coherent extrapolated volition of humankind*.

In poetic terms, our *coherent extrapolated volition* is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.

Knew more: Fred may believe that box A contains the diamond, and say, "I want box A." Actually, box B contains the diamond, and if Fred knew this fact, he would predictably say, "I want box B."

If Fred would adamantly refuse to even consider the possibility that box B contains a diamond, while also adamantly refusing to discuss what should happen in the event that he is wrong in this sort of case, and yet Fred would still be indignant and bewildered on finding that box A is empty, Fred's volition on this problem is *muddled*.

Thought faster: Suppose that your current self wants to use an elaborate system of ropes and sticks to obtain a tasty banana, but if you spent an extra week thinking about the problem, you would predictably see, and prefer, a simple and elegant way to get the banana using only three ropes and a teddy bear.

I arbitrarily declare the poetic term "think faster" to also cover *thinking smarter*, generalizing to straightforward transforms of existing cognitive processes to use more computing power, more neurons, et cetera. But the less understandable your present self finds your extrapolated volition, the greater the *distance*.

Were more the people we wished we were: Any given human is inconsistent under reflection. We all have parts of ourselves that we would change if we had the choice, whether minor or major.

Suppose Fred has a deeply repressed desire to murder Steve, but Fred consciously holds that murder is wrong. I do not call it "helping Fred," to murder Steve on his behalf. (Obviously I would not murder Steve regardless. I am an altruist, but also a private citizen; I need not help people in ways that conflict with my own ethics. I raise this scenario to ask whether murdering Steve is "helping Fred," considered in isolation.)

Suppose Fred *decides* to murder Steve, but when questioned, Fred says this is because Steve hurts other people, and needs to be stopped. Let's do something humans can't do, and peek inside Fred's mind-state. We find that Fred holds the verbal moral belief that hatred is never an appropriate reason to kill, and Fred hopes to someday grow into a celestial being of pure energy who won't hate anyone. We extrapolate other aspects of Fred's psychological growth, and find that this desire is expected to deepen and grow stronger over years, even after Fred realizes that the Islets worldview of "celestial beings of pure energy" is a myth. We also look at the history of Fred's mind-state and discover that Fred wants to kill Steve because Fred hates Steve's guts, and the rest is rationalization; extrapolating the result of diminishing Fred's hatred, we find that Fred would repudiate his desire to kill Steve, and be horrified at his earlier self.

I would construe Fred's volition not to include Fred's decision to kill Steve, noting that this is a *medium-distance* volition.

If self-modification were a routine part of human-existence, I suspect we would regard Fred's *predictable regret* after self-modifying as analogous to Fred's predictable regret after learning that box B contains the diamond. (That is, we would regard it as a short-distance extrapolation.)

We may want things we don't want to want. We may want things we wouldn't want to want if we knew more, thought faster. We may prefer not to have our extrapolated volition do things, in our name, which our future selves will predictably regret. The volitional dynamic takes this into account in multiple ways, including extrapolating our wish to be better people.

I distinguish "human," that which we are, from "humane"—that which, being human, we wish we were. Our extrapolated volition needs to encapsulate our aspiration to humaneness, not just our humanity.

Extrapolating Fred's wish to be a better person can amplify, or reduce, Fred's *spread*. If you are uncertain about a fundamental decision as to what kind of person you would like to be, this uncertainty may affect hundreds or millions of other decisions. All your future might hang in the balance, everything depending on who you decide to be when you grow up. Taking into account our wish to be better people can also reduce the *spread* in the extrapolation. Maybe you're presently uncertain whether to live in Detroit or Salt Lake City, but if you were a nicer person you'd predictably move to Australia. (No insult is intended to Australia.)

Had grown up farther together: A model of *humankind's coherent* extrapolated volition should not extrapolate the person you'd become if you made your decisions alone in a padded cell. Part of our predictable existence is that we predictably interact with other people. A dynamic for CEV must take a shot at extrapolating human interactions, not just so that the extrapolation is closer to reality, but so that the extrapolation can encapsulate memetic and social forces contributing to niceness.

Our CEV may judge some memetic dynamics as not worth extrapolating—not search out the most appealing trash-talk TV show.

Social interaction is probably intractable for real-world prediction, but no more so than individual volition. That is why I speak of *predictable* extrapolations, and of calculating the spread.

Where the extrapolation converges rather than diverges: Some of humanity's possible futures may hinge on decisions humanity has not yet made. As we grow up farther and make more decisions, our coherent extrapolated volition may change. Where the course of humankind is not presently predictable (exhibits significant *spread*), our CEV should take this into account, and leave options open, waiting for our decision.

Where our wishes cohere rather than interfere: Coherence is not a simple question of a majority vote. Coherence will reflect the balance, concentration, and strength of individual volitions. A minor, muddled preference of 60% of humanity might be countered by a strong, unmuddled preference of 10% of humanity. The variables are quantitative, not qualitative.

Extrapolated as we wish that extrapolated: Your volition should take into account only those desires and aspects of your personality that you would want your volition to take into account—the volitional dynamic should be consistent under reflection. This is a lesser special case of the rule that the Friendly AI should be consistent under reflection (which might involve the Friendly AI replacing itself with something else entirely). Deciding what your volition should take into account is treated separately because it comes earlier in order of evaluation. The dynamic would ask "Does Fred want this desire included in his volition?" before asking "What Friendly AI does Fred's extrapolated volition want?"

Interpreted as we wish that interpreted: I arbitrarily declare this part of the poetry to stand for the volitional dynamic's ability to renormalize, rewrite itself, or replace itself entirely.

3.1. Coherence and Influence

Coherence: Strong agreement between many extrapolated individual volitions which are unmuddled and unspread in the domain of agreement, and not countered by strong disagreement. Coherence

- increases, as more humans actively agree;
- decreases, as more humans actively disagree (the strength of opposition decreases if the opposition is muddled); and
- increases, as individuals support their wishes more, with stronger emotions or more settled philosophy.

It should be easier to counter coherence than to create coherence.

Influence: Acts carried out, optimization pressures applied, transformations undertaken, in the service of goals on which our extrapolated volitions exhibit coherence.

Less influence is required to *rule out* narrow slices of our possible futures, or prevent things from happening; more influence is required to direct humanity *into* specific narrow slices of future, or make specific things happen. The narrower the slice that our CEV wants to avoid, the less consensus required, the more distant the future selves that might produce the warning. The narrower the slice of the future that our CEV wants to actively steer humanity *into*, the *more* consensus required.

Distance, muddle, and spread will more rapidly attenuate positive influence than negative influence; the initial dynamic for CEV should be conservative about saying "yes," and listen carefully for "no."

Extrapolating your wish to be a better person may add considerable *distance*. If we extrapolate out far enough, we may end up with a Power, or something else too powerful and alien for your present-day self to comprehend. If our unimaginably distant future selves have an 80% probability of attaching a huge value to cheesecake for no reason our current selves can comprehend, this may not be a good reason to actively encourage present-day humans to fill their lives with cheesecake. It probably *is* a good reason to prevent people from *destroying* present-day cheesecakes. *Distance* and *spread* should attenuate the force of "do this" much more rapidly than they attenuate the force of "Yikes! Don't do that!" In qualitative terms, our unimaginably alien, powerful, and humane future selves should have a strong ability to say "Wait! Stop! You're going to predictably regret that!", but we should require much higher standards of predictability and coherence before we trust the extrapolation that says "Do this specific positive thing, even if you can't comprehend why."

A key point in building a *young* Friendly AI is that when the *chaos* in the system grows too high (spread and muddle both add to *chaos*), the Friendly AI does not *guess*. The young FAI leaves the problem pending and calls a programmer, or suspends, or undergoes a deterministic controlled shutdown. If humanity's volition is just too chaotic to extrapolate, the attempt to manifest our coherent extrapolated volition must fail *visibly* and *safely*.

3.2. Renormalizing the Dynamic

The FAI programmers do not need to get the dynamic exactly right the first time. They need to try their best to be perfect; but they do not need to *be* perfect.

I hope to construe a sufficiently powerful *initial dynamic* that I can ask: "What dynamic for extrapolating a volition would we *want*?" Or more generally, ask: "What should the code of a Friendly AI look like?" Or more generally still, "Write a piece of code that humanity wants to run." Perhaps the code humanity would prefer is not a dynamic for extrapolating a volition, or any kind of Friendly AI.

One cannot bootstrap from a vacuum. To ask what humanity wants requires an *initial dynamic* which is well-defined in an engineering sense, and which works well enough to answer extremely complex questions. It takes a *satisfactory* initial dynamic of volition to extrapolate an *optimal* dynamic of volition.

The task of construing a satisfactory initial dynamic is not so impossible as it seems. The satisfactory initial dynamic can be coded and tinkered with over years, and may improve itself in obvious and straightforward ways before taking on the task of rewriting itself entirely.

There is an obvious analogy between construing a satisfactory initial dynamic for extrapolating a volition, and developing an Artificial Intelligence smart enough to improve its own code and bootstrap to superintelligence. Friendly AI theory treats the second problem as a special case of the first.

3.3. Coherent Extrapolated Volition Is an Initial Dynamic

Once we have something that approximates a volition of the human species, that volition then has the chance to write its own superintelligence, optimization process, legislative procedure, god, or constitution. I try not to get caught up on CEV as a model of the *actual future*, even though it seems like a Nice Place To Live. The purpose of CEV as an initial dynamic is not to be *the* solution, but to ask what solution we want.

Nonetheless, CEV seems to me like it would function reasonably well as a Nice Place To Live, given some auxiliary dynamics—for example:

- Our coherent extrapolated volition should choose a small set of universal background rules, not optimize every detail of individual lives. (Rules in the sense of apparent behaviors of Nature, not legal regulations. An example of a *current* background rule is "As time goes on, you will age, become weaker and sicker, slowly lose neurons, and die." This strikes me as the diametric opposite of what I would prefer.)
- The rules should not be inscrutable, and should prefer to operate scrutably, so that humans can understand, predict, and manipulate their world.
- The ruleset should be small enough that human beings can learn and understand all the rules in reasonable time.

This would make our post-Singularity future *more* humanly comprehensible than any modern-day legal system.

Auxiliary dynamics of a Nice Place to Live should not be included in the *initial dynamic* because:

- They are complex, adding implementation difficulty.
- They are guesses, adding probability of error.
- They are means to an end.
- They are only relevant if humanity wants to live inside humanity's volition, rather than somewhere else.
- A decent initial dynamic must be capable (design requirement) of extrapolating auxiliary dynamics.
- Auxiliary dynamics often have a one-way aspect; once they're built into the structure, it's harder to renormalize them out than to renormalize them in.

- They are too much fun to speculate about.

There is a principled distinction between discussing CEV as an initial dynamic of Friendliness, and discussing CEV as a Nice Place to Live. Despite the apparent similarity, these are fundamentally different purposes, with different design requirements. When I originally wrote this essay, I was conflating the two purposes. Yes! This essay is *already obsolete*! There's not one single damn thing I've tried to write in the last three years that was still current with my thinking when I published it, and in this case it only took *two days*!

Ahem. Okay, let's see what needs changing from what I've written so far . . . The rules about degree of influence depending on the balance of coherence might need to allow a sufficiently strong win to rewrite the initial dynamic entirely, rather than striking something resembling a political compromise. Otherwise there's too much inertia in the system—it's too hard to get sufficient agreement to make large leaps forward from the programmers' initial solution. I think I would wish to live in a world that required compromises when writing rules, but not meta-rules. Compromise is itself a meta-rule, and it's almost impossible to get rid of compromise if you have to compromise on how to get rid of it. Initial dynamics should not start out having already taken one-way steps. Thus, I now regard the previously offered rules about coherence and influence as auxiliary dynamics of a Nice Place to Live; they might make an appearance in the initial dynamic, but only if that doesn't create inertial mud in the system, or require unrealistic unanimity to leap to better meta-rules, or otherwise make it difficult to write a coherent dynamic on the next round of renormalization.

(I left this in the final version of the paper to emphasize that I am still working out these issues. For all I know, I may scrap coherent extrapolated volition entirely in 2005.)

4. Caring about Volition

Some have asked: "Why care about our extrapolated volitions? Like, at all?"

The dynamic of extrapolated volition refracts through that cognitive complexity of human minds which lead us to care about all the other things we might want; love, laughter, life, fairness, fun, sociality, self-reliance, morality, naughtiness, and anything else we might treasure. Whatever you care about—whatever it is of which you say, "XYZ is what we should care about"—it's your brain, a human brain, that produced that statement; and the air vibrating on your lips exists in a universe of cause and effect. There is a physical explanation for why you said the words "XYZ." This does *not* devalue XYZ. There is a complex psychological tangle here, which leads people to experience feelings of existential emptiness for no good reason. Let it stand that human statements about morality are explicable in terms of cause-and-effect physics, *just like absolutely everything* *else,* and this in no way detracts from the force of the moral statements. Human moral psychology is not *merely* physics; physics is not "mere." But everything that exists, exists embedded within physics.

CEV is a dynamic that passes through, and extrapolates, the human psychology of decision-making—including argument about morality, and human emotion, and the memes we argue together; everything that changes our choices. Any thoughts you have about morality are included as a special case of CEV extrapolating your thoughts.

If you think that some particular morality should be, not a special case of your extrapolated volition, but the law and the whole of the law, I can only lift my hands helplessly, and say that I have no time to hear out your argument from the other million arguments I might hear. If there is some chain of moral logic so self-evident that any human who hears with open mind must follow, then that is what the *initial dynamic* of extrapolated volition would produce on the next round. I don't think this is what will happen, but I took into account the possibility in planning.

If the Pnakotic Manuscripts are truly the source of all goodness and light, you're covered.

If humans make their judgments using hidden superpowers of carbon atoms that no mere silicon atoms can match, the attempt to construct a coherent extrapolated volition will fail early, visibly, and harmlessly when the dynamic finds it can't predict human moral judgments. Likewise if the dynamic learns to predict humans successfully, and then predicts that humans wouldn't want to be predicted by a mere machine without an immortal soul.

These special cases shouldn't require special code. They're just implicit in a volitional dynamic that does the Right Thing, the thing *we want* to happen.

I didn't choose coherent extrapolated volition lightly; I chose it as the endpoint of a journey that began by asking "What should I care about?" I don't know all of the answer, but at least I now understand the nature of the question; it is no longer mysterious to me. As for the answer, I have encountered many subproblems which had surprising and compelling resolutions. But I couldn't predict the solution or even the *kind* of solution in advance, nor guess which part of the problem I would see in a new light. I do not know the library of forces which compel me. I do not know the search space of arguments which move me. I do not accept new parts of myself blindly, when I discover them; I am compelled to judge my compulsions, moved to question my movements. Yet there may be light within me that I do not realize, but would acknowledge if shown. All I can do is take that reply which I would give if I knew more, thought faster, were more the person I wished I were, had grown up farther with you all, and call it the right answer. But I cannot tell you specifically the right answer. In the end, it reduces to me chipping away at the question, waiting to be struck by a good idea.

That chipping process is what I intend CEV to extrapolate. If there is a grand unifying theory waiting to strike, and it is the same for everyone, then chipping away is the only method I currently know for finding it. As I understand the question it is quite unlikely that there is a grand unifying theory, and so we shall have to find the answer the hard way, by making the decision. We cannot ask our volitions for an answer, if that is the *only* answering procedure we know.

4.1. Motivations

Defend humans, the future of humankind, and humane nature.

These are three things which the wrong choice of initial dynamic might screw up. Most people careless enough to make blatantly foolish choices for AI morality are also foolish enough to screw up the technical implementation. If someone programs an AI using pictures of smiling humans as reinforcement, the resulting AI will *not* keep everyone happy whether they like it or not; the resulting AI will tile the solar system with tiny molecular smiley faces. Mostly, the meddling dabblers won't trap you in *With Folded Hands* or *The Metamorphosis of Prime Intellect*. Mostly, they're just gonna kill ya. It's unlikely that any given meddling dabbler will succeed; but there's a plentiful supply of dabblers, and Moore's Law keeps lowering the bar.

But if we imagine the improbable event of a meddling dabbler somehow succeeding in solving the deep technical problems of Friendly AI, yet not thinking through Friendliness, then we can imagine scenarios such as *With Folded Hands*—where the robots protect human life, and prevent humans from experiencing pain or distress, but care nothing for those other things that humans love, such as liberty. The future of humankind is lost.

I value my friends and fellows: the six billions now in jeopardy, their lives and liberties.

I value the future of humankind: trillions or quadrillions or octillions of lives; stars and galaxies become our cities. Weightier by far than a mere six billion, it might seem; but we six billion carry the seed complexity. As best I can foresee, at least some of the key decisions lie rightfully in our hands; not beyond the rethinking of future generations, but ourselves creating and becoming the minds who rethink. It's one heck of a huge future hanging in the balance, and I really, really, really don't want to spoil it.

I value the light in the heart of humanity: love, laughter, life, happiness, fun, fairness, compassion, empathy, sympathy, sociality, sexuality, self-reliance, shared warmth and togetherness, moral argument, naughtiness, serene wisdom, juvenile mischief; our wish to be better people; humane nature, and the future our humane drives will shape.

The initial dynamic of CEV takes into account the wishes of those billions now in existence, including our wish to live, and our wish for freedom, and our wish to defend those other things we value.

CEV looks forward in time, to find if our past predictably threatens portions of our future we will predictably value. Since the output of the CEV *is* one of the major forces shaping the future, I'm still pondering the order-of-evaluation problem to prevent this from becoming an infinite recursion.

Extrapolating volition seeks out the light in human nature. I think. I'm not sure about this part. (Other sections touch on this in more detail.)

Encapsulate moral growth.

Humanity's moral memes have improved greatly over the last few thousand years. How much farther do we have left to go? I was born too late myself, but you don't need to be very old to *remember* a time when blacks rode in the back of the bus—a thought that still chills me, that the time is so close to our own. What if we too look like barbarians from the perspective of a few years down the road?

We need to extrapolate our moral growth far enough into the virtual future to discover our gasp of horrified realization when we realize that the common modern practice of X is a crime against humanity, even though, like slavery, X seemed to us like a normal and harmless institution. Possible values of X include sending children to school, keeping chimpanzees in cages, not knowing the basic laws of physics, buying mass-manufactured tools, or living without access to sandestins.

(This was phrased as a Nice Place to Live argument, but it also applies to the initial dynamic.)

Humankind should not spend the rest of eternity desperately wishing that the programmers had done something differently.

This seems obvious, until you realize that only the Singularity Institute has even *tried* to address this issue. I haven't seen a single other proposal for AI morality out there, not even a casual guess, that takes the possibility of Singularity Regret into account. Not one. Everyone has their brilliant idea for the Four Great Moral Principles That Are All We Need To Program Into AIs, and not one says, "But wait, what if I got the Four Great Moral Principles *wrong*?"

They don't think of writing any escape clause, any emergency exit if the programmers made the wrong decision. They don't wonder if the original programmers of the AI might not be the wisest members of the human species; or if even the wisest human-level minds might flunk the test; or if humankind might outgrow the programmers' brilliantly insightful moral philosophy a few million years hence.

But most who ponder AI morality walk straight into the whirling razor blades, without the slightest hint of fear. I am dismayed, but not surprised. I do *not* say it is inevitable; someone who rises to the challenge can do better than that. I never made that error, not once. Once upon a time, back at the age of sixteen, I made one hell of a huge philosophical blunder about how morality worked, which caused me to think that to create a Friendly AI I just needed to create a really powerful self-improving optimization process and wave it cheerily on its way. But I *never* thought I had the wisdom or authority to lay down Ten Commandments, not for an AI and not for humanity. The problem was so uncomfortable that it took me until the age of twenty to acknowledge the problem existed. (Youthful foolishness I don't intend to ever repeat; I think I've changed sufficiently since then.)

Once I acknowledged the problem existed, I didn't waste time planning the New World Order. No, not even a fraction of a second; it was too obvious a failure mode. The first thing I said was, "What if the programmers don't get it right on the first try?", followed immediately by, "A self-modifying expected utility maximizer with a constant utility function won't let anyone tamper with its utility function; I need to figure out how to construct a dynamic for a correctable utility function such that the dynamic is stable under self-modification." I didn't use those words. I knew the math, but I didn't know the standard names. I made up silly names of my own, because I was young. But I wasn't young enough to rub my hands in glee at the prospect of laying down the Four Great Moral Principles. I wasn't even young enough to be frightened at the thought of making moral choices for the whole human species. I took for *granted* that whichever shmuck happened to program the first superintelligence had no special moral privileges, and I looked for ways to refer the question back to humanity. I looked to Friendliness architecture and not Friendliness content, because I knew the difference between fun moral argument and technical thinking. I knew which kind of thinking was useful, and which kind of thinking was not. I flinched away from addressing the problem of Friendliness content, and took refuge in Friendly AI structure. And whaddaya know, it *worked*. Once I had a language to describe solutions, it became far easier to come up with a solution. That younger self didn't get the theory right on his first try, not even close. But I can see the foundations of the extrapolated volition model, even in a younger self so distant that I sometimes can't guess what he was thinking.

The extrapolated volition model doesn't have an escape hatch *tacked on*—the escape hatch is not an awkward added clause, an Eleventh Commandment. The escape hatch is implicit in extrapolated volition doing the Right Thing. If our extrapolated volitions say we don't want our extrapolated volitions manifested, the system replaces itself with something else we want, or vanishes in a puff of smoke (undergoes an orderly shutdown). And our CEV uses greater intelligence to search for errors, unforeseen consequences, future horrors, the predictable moral revulsions of tomorrow's selves. If we scream, the rules change; if we predictably scream later, the rules change now.

It may be hard to get CEV right—come up with an AI dynamic such that our volition, as defined, is what we intuitively *want*. The technical challenge may be too hard; the problems I'm still working out may be impossible or ill-defined. I don't intend to trust any design until I see that it works, and only to the extent I see that it works. Intentions are not always realized. But I *intend* that CEV have an intrinsic escape hatch, and search for unforeseen consequences, and not rely on the ability of the original programmers to dictate all the moral questions correctly on their first try. The other plans I hear proposed would fail even if they succeeded.

There are fundamental reasons why Four Great Moral Principles or Ten Commandments or Three Laws of Robotics are wrong as a design principle. It is anthropomorphism. One cannot build a mind from scratch with the same lofty moral statements used to argue philosophy with pre-existing human minds. The same people who aren't frightened by the prospect of making moral decisions for the whole human species lack the interdisciplinary background to know how much complexity there is in human psychology, and why our shared emotional psychology is an invisible background assumption in human interactions, and why their Ten Commandments only make sense if you're already a human. They imagine the effect their Ten Commandments would produce upon an attentive human student, and then suppose that telling their Ten Commandments to an AI would produce the same effect. Even if this worked, it would still be a bad idea; you'd lose everything that wasn't in the Ten Commandments.

Think of a utility function, $U(x)$, with additive components $U_1(x)$, $U_2(x)$, $U_3(x)$, and so on. There are hundreds of drives in human psychology. Even if Ten Commandments worked to transfer rough analogues of the first ten components, $V_1(x)$ through $V_{10}(x)$, you'd still have disagreements because of the sloppy transfer method, places where V_1 didn't match U_1. Human psychology doesn't run on utility functions; just the choice of formalism implies some degree of mismatch with actual decisions. But forget the minor mismatch; what about $U_{11}(x)$ through $U_{245}(x)$? What about everything that's not in the Ten Commandments? The result would be humanity screaming as the AI took actions in total disregard of $U_{11}(x)$ and all other components the original programmers got wrong or didn't consider—an AI that kept us from harm at the cost of our liberty, and so on.

Usually the Ten Commandments folks don't even advocate transferring ten underlying human emotions $U_1(x)$ through $U_{10}(x)$—just trying to transfer ten specific moral arguments that sound good to their human psychologies, without transferring any of the underlying human emotions to lend a commonsense interpretation. Humanity would scream wherever the literal extrapolation of those ten statements conflicted with one of our hundreds of psychological drives, or with one of the thousands of moral statements not included on the list of ten. Not that this is a likely actual scenario. Anyone ignorant enough to make such a proposal is ignorant enough to kill you outright if they succeed, build their AI that Pays Attention To Social Entities and goes on to tile the solar system with tiny interacting agents pushing around billiard balls.

Avoid hijacking the destiny of humankind.

FAI programmers are ordinary shmucks and do not deserve, a priori, to cast a vote larger than anyone else. There is a challenge of the professional ethics of FAI, as there is the challenge of Friendliness, and the challenge of FAI theory, and the challenge of saving the world without distraction. Avoid the obvious traps against which history books and evolutionary psychology warn you; act as you would wish another to act in your shoes; discharge your duties without exerting undue influence. It seems obvious enough, yet when I dare to say it out loud . . . I'd get less flack if I said I wanted to transport the entire human species into an alternate dimension based on a randomly selected hentai anime. Certain people are *terrified* at the thought that anyone might lay claim to being fair, as if that were an excuse for not trying.

"But what if Dennis calls it fairness that he should own the world outright, and objects to this notion of coherent extrapolated volition as unfair?" Even if you are utterly selfish, the answer is still: "If I cannot tell Eliezer my own name in Dennis's place, the best general principle I can tell Eliezer is that he should not listen to such arguments." I have not yet heard the objectors protest in their own right, rather than on Dennis's behalf. For if they proposed a concrete alternative, put forward their own desire rather than a hypothesis of how someone *else* might protest, they should themselves need to confront objections. And that is not a philosopher's place, in the scheme of things.

The initial conditions of the AI have a large *de facto* effect on the human species. Even if our coherent extrapolated volition wants something other than a CEV, the programmers choose the starting point of this renormalization process; they must construct a *satisfactory* definition of volition to extrapolate an improved or optimal definition of volition.

Different classes of satisfactory initial definitions may fall into different self-consistent attractors for optimal definitions of volition. Or they may all converge to essentially the same endpoint. A CEV might survey the "space" of initial dynamics and self-consistent final dynamics, looking to see if one alternative obviously stands out as best; extrapolating the opinions humane philosophers might have of that space. But if there are multiple, self-consistent, satisficing endpoints, each of them optimal under their own criterion—okay. Whatever. As long as we end up in a Nice Place to Live.

And yes, the programmers' choices may have a huge impact on the ultimate destiny of the human species. Or a bird, chirping in the programmers' window. Or a science fiction novel, or a few lines spoken by a character in an anime, or a webcomic. Life is chaotic, small things have large effects. So it goes. I started thinking about volitional models of Friendliness when I read Greg Egan's *Diaspora*; blame the next billion years on him. Not that I think humankind will use CEV forever. But who knows how much humankind's future may be shaped by our early days?

Even if the programmers are nice people and ground the AI in all humanity, the choice of grounding mechanism makes a difference. The decision to hand decisions to humanity is still a decision.

Whether the programmers like it or not, the programmers' choices will make huge differences, and not just because of the butterfly effect. *But there is no need to be a jerk about it.* A determined altruist can always find a way to cooperate on the Prisoner's Dilemma. We are not talking about programmers imposing their Ten Commandments on the rest of humanity until the end of time. We are talking about people stuck with a job that is frankly absurd, doing their best not to be jerks. The dynamics that extrapolate volition won't resemble fun political arguments that humans enjoy. The dynamics will be choices of mathematical viewpoint, computer programs, optimization targets, reinforcement criteria, and AI training games with teams of agents manipulating billiard balls. The initial dynamic won't mention Democrats or Republicans, or whether Germany or France gets the Alsace-Lorraines, or who becomes ruler of Australia. The initial dynamic may renormalize to something entirely different, and probably will.

Maybe, depending on the random seeds of the AI program, *you* could end up as ruler of Australia! But which random seed? `0xFA34E1E8`? `0x81B2AA09`? I don't think there's much point in fighting. I'll roll thirteen six-sided dice to determine the randseeds, and then auction them off on eBay as THE DICE THAT DETERMINED THE FATE OF THE HUMAN SPECIES. And if there's any particular value for a randseed that people can *agree* they want, well, that's what we have a CEV to renormalize to, right?

Avoid creating a motive for modern-day humans to fight over the initial dynamic.

One of the occasional questions I get asked is "What if al-Qaeda programmers write an AI?" I am not quite sure how this constitutes an *objection* to the Singularity Institute's work, but the answer is that the solar system would be tiled with tiny copies of the Qur'an. Needless to say, this is *much* more worrisome than the solar system being tiled with tiny copies of smiley faces or reward buttons. I'll worry about terrorists writing AIs when I am through worrying about brilliant young well-intentioned university AI researchers with millions of dollars in venture capital. *The outcome is exactly the same,* and the academic and corporate researchers are far more likely to do it first. This is a *critical* point to keep in mind, as otherwise it provides an excuse to go back to screaming about politics, which feels so much more satisfying. When you scream about politics you are really making progress, according to an evolved psychology that thinks you are in a hunter-gatherer tribe of two hundred people. To save the human species you must first ignore a hundred tempting distractions.

I think the objection is that, in theory, someone can disagree about what a superintelligence ought to do. Like Dennis, who thinks he ought to own the world outright. But do you, as a third party, want me to pay attention to Dennis? You can't advise me to hand the world to *you*, personally; I'll delete your name from any advice you give me before I look at it. So if you're not allowed to mention your own name, what general policy do you want me to follow?

Let's suppose that the al-Qaeda programmers are brilliant enough to have a realistic chance of not only creating humanity's first Artificial Intelligence but also solving the technical side of the FAI problem. Humanity is not automatically screwed. We're postulating some *extraordinary* terrorists. They didn't fall off the first cliff they encountered on the technical side of Friendly AI. They are cautious enough and scared enough to double-check themselves. They are rational enough to avoid tempting fallacies, and extract themselves from mistakes of the existing literature. The al-Qaeda programmers will not set down Four Great Moral Principles, not if they have enough intelligence to solve the technical problems of Friendly AI. The terrorists have studied evolutionary psychology and Bayesian decision theory and many other sciences. If we postulate such extraordinary terrorists, perhaps we can go one step further, and postulate terrorists with moral caution, and a sense of historical perspective? We will assume that the terrorists still have all the standard al-Qaeda morals; they would reduce Israel and the United States to ash, they would subordinate women to men. Still, is humankind screwed?

Let us suppose that the al-Qaeda programmers possess a deep personal fear of screwing up humankind's bright future, in which Islam conquers the United States and then spreads across stars and galaxies. The terrorists know they are not wise. They do not know that they are evil, remorseless, stupid terrorists, the incarnation of All That Is Bad; people like *that* live in the United States. They are nice people, by their lights. They have enough caution not to simply fall off the first cliff in Friendly AI. They don't want to screw up the future of Islam, or hear future Muslim scholars scream in horror on contemplating their AI. So they try to set down precautions and safeguards, to keep themselves from screwing up.

One day, one of the terrorist programmers says: "Here's an interesting thought experiment. Suppose there were an atheistic American Jew, writing a superintelligence; what advice would we give him, to make sure that even one so steeped in wickedness does not ruin the future of Islam? Let us follow that advice ourselves, for we too are sinners." And another terrorist on the project team says: "Tell him to study the holy Qur'an, and diligently implement what is found there." And another says: "It was specified that he was an atheistic American Jew, he'd never take that advice. The point of the thought experiment is to search for general heuristics strong enough to leap out of really fundamental errors, the errors we're making ourselves, but don't know about. What if he should interpret the Qur'an wrongly?" And another says: "If we find any truly general advice, the argument to persuade the atheistic American Jew to accept it would be to point out that it is the same advice he would want *us* to follow." And another says: "But he is a member of the Great Satan; he would only write an AI that would crush Islam." And another says: "We necessarily postulate an atheistic Jew of exceptional caution and rationality, as otherwise his AI would tile the solar system with American music videos. I know no one like that would be an atheistic Jew, but try to follow the thought experiment."

I ask myself what advice I would give to terrorists, if they were programming a superintelligence and honestly wanted not to screw it up, and then that is the advice I follow myself.

The terrorists, I think, would advise me not to trust the self of this passing moment, but try to extrapolate an Eliezer who knew more, thought faster, were more the person I wished I were, had grown up farther together with humanity. Such an Eliezer might be able to leap out of his fundamental errors. And the terrorists, still fearing that I bore too deeply the stamp of my mistakes, would advise me to include all the world in my extrapolation, being unable to advise me to include only Islam.

But perhaps the terrorists are still worried; after all, only a quarter of the world is Islamic. So they would advise me to extrapolate out to *medium-distance*, even against the force of muddled short-distance opposition, far enough to reach (they think) the coherence of all seeing the light of Islam. What about extrapolating out to long-distance volitions? I think the terrorists and I would look at each other, and shrug helplessly, and leave it up to our medium-distance volitions to decide. I can see turning the world over to an incomprehensible volition, but I would want there to be a comprehensible reason. Otherwise it is hard for me to remember why I care.

Suppose we filter out all the AI projects run by Dennises who just want to take over the world, and all the AI projects without the moral caution to fear themselves flawed, leaving only those AI projects that would prefer *not* to create a motive for present-day humans to fight over the initial conditions of the AI. Do these *remaining* AI projects have anything to fight over? This is an interesting question, and I honestly don't know. In the real world there are currently only a handful of AI projects that might dabble. To the best of my knowledge, there isn't more than one project that rises to the challenge of moral caution, let alone rises to the challenge of FAI theory, so I don't know if two such projects would find themselves unable to agree. I think we would probably agree that we didn't know whether we had anything to fight over, and as long as we didn't know, we could agree not to care. A determined altruist can always find a way to cooperate on the Prisoner's Dilemma.

Keep humankind ultimately in charge of its own destiny.

Coherent extrapolated volition appears to me to embody an intrinsic switchover to a direct voting system when people are grown enough to handle it. Your immediate decision, the whim of this present moment, does count for something in your extrapolated volition. But if your future self would predictably shriek in horror, that extrapolation may win out, despite the added distance. When people know enough, are smart enough, experienced enough, wise enough, that their volitions are not so incoherent with their decisions, their direct vote could determine their volition. If you look closely at the reason why direct voting is a bad idea, it's that people's decisions are incoherent with their volitions. (Again, note the elegance; the slow growth into direct voting is not a tacked-on special case, just the dynamic doing the Right Thing.)

I view it as a philosophical advantage of extrapolated volition that at no point does humanity turn its destiny over to an external god, even a humane external god. (But better that than being dead. Applied Theology is still a definite option if extrapolated volition doesn't work out.)

And no, a Really Powerful Optimization Process is not a god.

Help people.

I want to leave the world a better place than I found it. The intention behind this project of manifesting the coherent extrapolated volition of humankind is to actually help humans; as opposed to, say, harming them.

4.2. But What If This Volition Thing Doesn't Work?

What if the initial dynamic works as *designed*, but not as *planned*, harming people instead of helping them?

This is an objection that applies to *any* initial dynamic, not just extrapolated volition: "What if the dynamic just doesn't work out the way you thought it would?" The programmers are under obligation not to be jerks, to craft a reasonable and satisfactory initial dynamic to the best of their abilities, without playing nitwit games. But honest goodness does not guarantee a good outcome, if vision fails—that event which we who are wise in the art of caution call a "mistake." (Some who know not the art have heard that other folks make "mistakes," but they think it a moral parable to teach humility, and so they respond by abasing themselves and acknowledging their theoretical fallibility, then continue doing whatever they planned to do anyway. Rather than, say, devising safeguards and recovery plans.)

Can the programmers trust their choice of initial dynamic so highly? Maybe they should peek at the extrapolated outcome, make sure that everything works out all right, before they set their creation irrevocably in motion.

But should the past have a veto over the future?

Consider the horror of America in 1800, faced with America in 2000. The abolitionists might be glad that slavery had been abolished. Others might be horrified, seeing federal law forcing upon all states a few whites' personal opinions on the philosophical question of whether blacks were people, rather than the whites in each state voting for themselves. Even most abolitionists would recoil from in disgust from interracial marriages—questioning, perhaps, if the abolition of slavery were a good idea, if this were where it led. Imagine someone from 1800 viewing *The Matrix*, or watching scantily clad dancers on MTV. I've seen movies made in the 1950s, and I've been struck at how the characters are *different*—stranger than most of the extraterrestrials, and AIs, I've seen in the movies of our own age. Aliens from the past.

Something about humanity's post-Singularity future will horrify us. It is guaranteed, no matter how good things get. It is guaranteed *because* things will get better. Imagine the culture shock if the eighteenth century got a look at the twentieth century. What about the tenth century? And that's not much of a gap; everyone, then and now, was human.

Maybe 6-year-olds will be considered above the age of consent. Maybe humanity will abandon private property. Maybe rape will be legal. Maybe everyone will be uploaded whether they like it or not. Maybe everyone wearing a Japanese schoolgirl uniform at the time of Singularity will be attacked by tentacle monsters. *Something* that would shock the people of this present day. I am reminded of a book I once read, a cyberpunk role-playing game set in the middle of the twenty-first century, which in the midst of describing body replacements and organ rippers, mind-control drugs and corporate assassinations, mentioned how a twenty-second century Solid Citizen might do things that would horrify, disgust, and *frighten* that twenty-first century netrunner.

It is not trivial, to suggest that the programmers dare to peek at the outcome. Or that the programmers appoint a Last Judge to check, someone chosen for their tolerance and broad-mindedness as the last safeguard of the human species, to peek at the outcome and return one bit of information, 1 or 0. Neither is it trivial to suggest that the programmers trust their initial dynamics *without* anyone peeking.

According to my current understanding of Friendly AI, there is a general safety rule which says, roughly, "Don't trust anything—dynamics, information, moral principles—until you see that it works better than you do." Subsystems that you trust may not be trusted to function together as supersystems. You wouldn't want to set an FAI dynamic irrevocably in motion (including an FAI dynamic that manifested the extrapolated volition of humankind) until you'd verified that the FAI's performance on the "checking an FAI works sensibly" problem was significantly better than the programmers' performance. I think, though I am not sure, that the FAI theory will dictate what constitutes a necessary and sufficient verification process, including whether we need a Last Judge to peek at the extrapolated final outcome.

Let it stand that the thought has occurred to me, and that I don't plan on blindly trusting *anything*. Recent FAI theory says that it shouldn't be necessary to blindly trust, but FAI theory might require a Last Judge, which is nontrivial to set up safely. If FAI theory requires a Last Judge, I would advise those al-Qaeda terrorists to set up a *lot* of safeguards. I don't want them to discover that we'll all end up as atheists; I want them to think of Islam triumphant, and not be tempted to meddle.

This problem deserves an essay in itself, which I may or may not have time to write.

Why not base the Friendly AI on individual volition instead of coherence in humankind's extrapolated volition?

Why not choose a Friendly AI that gives every individual their own wish? Why extrapolate a *collective* volition, and risk a tyranny of the extrapolated majority? If people want to be influenced by majority-coherence of extrapolated volitions, why not let their individual extrapolated volitions make that choice?

Individual volition would make sense as a final dynamic, but it can't be an *initial* dynamic. There's no way for individually extrapolated volitions to rewrite the entire process, redesign the meta-process that describes how to partition the individual volitions and how they interact. Any mistakes the programmers make in choosing the meta-process would be enshrined forever. You can go from a collective dynamic to an individual dynamic, but not the other way around; it's a one-way hatch. If you use an individual dynamic, that *forces* you to live in that particular world—which might not be the happiest world for us to live in. Maybe we could have more fun somewhere else.

How would you define what constitutes an "individual"? If, as I suggest, you let the coherence in the extrapolated volitions of a population produce a single final output decision; then the programmers base the initial dynamic off humanity's billions as they exist today, and let the majority volition choose whether to extend citizenship to chimpanzees on the next round of renormalization. If you enshrine forever the meta-rules for how individual volitions interact, the programmers have to get the meta-rules absolutely right, until the end of time, on the very first try.

Also, are you *sure* you don't want a tyranny of the 95% majority?

What about infants? What about brain-damage cases? What about people with Alzheimer's disease? If you extrapolated their pure individual volitions, you might find that they lacked the moral and social complexity to *want* to be rescued, as we of the 95% would define "rescue." A too-literal interpretation of individualist philosophy might wrench infants off their course to becoming humans and turn them into autistic super-infants instead. It's only genes and human parents who have this idea that infants are *destined* to become humans. It's not actually in infant *psychologies*, their mind-states at the age of six months.

(See also CalvinAndHobbes.)

This is a fundamental moral question, and the moral answer isn't obvious. Maybe the civilization of the thirty-second century would say: "Let them go their own way, what's wrong with that? There are far stranger minds, in these days, than super-infants; who are you to judge their choice?"

Or maybe humankind will define a spectrum of satisficing initial mind states, and require any mind not initially in a satisficing mind state to grow into a satisficing mind-state. In the sense that, for example, a 6-year-old would be required to grow into an 18-year-old before deciding where to go from there. Whether an older and wiser mind could voluntarily grow back into a 6-year-old psychology is an interesting question, and I think a distinct one.

Maybe there are people who care about individual rights so strongly that they don't care *what* a majority of extrapolated volitions say; they want to give infants and autistics their own private universes because that's what their philosophy says is right, and their philosophy makes no mention of what other people think of the issue.

Maybe future humanity will agree with them.

Maybe afterward everyone will say: "Eliezer was a bastard for endangering the free wills of infants that way, plotting to subject them to the tyranny of the extrapolated majority, daring even to think that one extrapolated volition should overrule another."

If so, the majority of extrapolated volitions will correct my folly, with no harm done. But to me, wrenching infants off their course to humanness seems like following philosophy off a cliff. Even if I thought otherwise, I would fear making a mistake. Choosing individual volition as an *initial* dynamic would *pre-emptively* make that moral choice; it is implicit in the *structure* of an individual dynamic. An individual dynamic would not take into account anything outside the infant unless something *in the infant's* extrapolated volition said so.

I think the extrapolated volitions of the 95% majority, including myself, may say that there are forces in our Nice Place to Live that do not derive strictly from individual volitions, that people can sometimes affect one another even without the consent of the affected. The *possibility* of occasional nonconsensual effect is a *deep* fundamental of human existence, not lightly to be thrown away. Would this mean that you needed to pay some attention to the course of humankind as a whole, lest the majority someday bite? Rather than holing up in a cave forever without fear? Maybe that isn't such a bad thing.

Or maybe the majority will fear the tyranny of the majority too greatly, rule out the smallest touch. I don't think it's my place to make the decision. If I later find I'm one of the 5% of humanity whose personal philosophies predictably work out to pure libertarianism, and I threw away my one chance to free humanity—the hell with it. Better than being dead. I don't see how this case differs from the other ways I might grow up to hold a different morality from the majority. Here and now, I don't know, and I can concentrate on not being a jerk.

Care to volunteer as Last Judge, so you can peek at the weeping and bereft parents, read the sage justifications in moral philosophy, and decide whether this constitutes a reason to shut down the AI? I who know much and think quickly do not expect that to happen, in exact proportion to how much I don't want it to happen. But if I wrote down ten such questions, I would not get ten correct answers. It's the sort of dilemma a Last Judge might face.

All those other darned humans in the world are no-good rotten bastards; why would their extrapolated volitions be any better, you blind optimist?

Let's suppose that all those darned other humans in the world are just no damn good. By comparison to what? The judgment implies that we must have some criterion in mind. Does the speaker have some better outcome in mind? It's not a logical implication—we could just be screwed. But people tend to imagine things they're afraid people want, or silly things for an extrapolated volition to do, and then look on the imagined outcome and say "Yuck." Extrapolated volition doesn't refract through the part of you that imagines the bad outcome; extrapolated volition refracts through the part of you that looks upon the imagined outcome, and says, "Yuck, I don't want that to happen." The ability to make this judgment exists in you, in your mind and brain. It has to come from humans, be generated by humans in some way. Same thing if you worry that humans are rotten bastards. Rotten bastards by comparison to what image that your mind generates? And how does a rotten bastard generate that image of something nicer? We need that "something nicer," and it has to be extracted from rotten bastards, and extrapolated volition is my guess how to do it.

Let's suppose, leaving aside the moral and technical questions, that someone took the initial dynamic and biased it toward positive emotions—extrapolated compassion, without extrapolating hatred. Easy enough to say, but whence came your decision that compassion was positive, and hate wasn't? *You* made the decision, and that's part of what gets extrapolated, under the heading of "our wish to be better people." Rather than directly trying to guess at which emotions are "positive" or "negative," I'm taking a step back, looking at my ability to make that judgment, and working it into the dynamic of extrapolated volition.

But suppose that only a few people wish they were better people, in a sense sufficient for extrapolated volition to look past the evil and the nastiness? It strikes me as unlikely, based on strictly intuitive sense data (note: this is worthless reasoning) that a majority of humanity have no yearnings in this direction. It looks culturally universal. But I could be wrong. The trouble is that half the people are below average, and their interviews don't often show up in the history books, and we don't hear as much about their memes. (Note that memetic forces generated by social interactions are also taken into account, if we want them to be, under the heading of "grown up farther together.")

The worrying question is: What if only 20% of the planetary population is nice, or cares about niceness, or *falls into the niceness attractor* when their volition is extrapolated? The notion being that this is enough to account for all the famous cases of niceness that everyone prefers to pass on and remember, and yet the majority of people on Earth are bastards. "Bastard volitions" will not vote to throw themselves into everlasting hell—ordinary selfishness being enough to prevent that—but they might create rules much to the distaste of us 20% (of course *you're* not a bastard, right?) For example, a rule stating that anyone who has computing power can do anything they want with that computing power, including creating innocent sentients and torturing them. It doesn't disadvantage any of the six billion people voting on it, except those silly 20% who care about fairness, so the vote passes. Similarly, the bastard volitions might vote to rewrite the initial dynamic to include no one except themselves. Maybe, if it's 80% of humanity in one coherent unit, and 20% in the other coherent unit, the 80% would vote to disenfranchise the 20%, or ensure that the 20% also grew into selfish bastards.

As I currently construe CEV, this is a real possibility. Trying to tinker with the initial dynamic to rule it out is dangerous, because of added complexity, and added probability of error, and too high a chance of ruling out something that made more sense than you thought it did. If one contemplates tampering with CEV just to make sure it comes out "right," one may as well abandon that solution entirely; what would be the point? (See Previously Asked Question 4, below, for an extended discussion.)

Or it might be that plain, honest CEV extrapolates far enough to reach a grown-up, kinder humankind. Wouldn't you be terribly ashamed to go down in history as having meddled with that, using lesser intelligence and lesser niceness, because you didn't trust your fellows? Wouldn't you be sad to have made a mockery of humanity's self-determination, if it would have worked anyway? Wouldn't you be horrified to find that you'd screwed it up?

I see the programmers as having an unalterable duty not to tinker with a running CEV; this constitutes taking over the world, which is one of many ways for seed AI programmers to be jerks. But the programmers are under no *obligation* to implement CEV, if it is something of which they don't want to be part. I'm an individual, with my own life and philosophy. There are many acts of which I am prohibited (because I say so, that's why). But I can always choose *not* to do something, even halt work already begun; that is my unalienable right. I can't tinker with a running CEV, but I can choose to dissipate the AI, or hold a vote of the programming team to dissipate the AI, or delegate a Last Judge to make that decision. Steering wheels are forbidden; off switches are not. I think there must come a time when the last decision is made and the AI set irrevocably in motion, with the programmers playing no further special role in the dynamics. But at any point before then, one can always say "No," and humanity will be no worse off than before.

Then what? Give up entirely? No, extrapolate from an idealized generic human, or an idealized generic human with an initial push toward altruism, or an idealized generic human with an initial push toward transpersonal philosophy, or extrapolate an imaginary civilization composed of genetically diverse individuals in the 99% niceness bracket . . .

If you started flipping through solutions that satisficed, looking for one that suited you alone *especially* well, I would call that gerrymandering. But I think the programmers should be allowed, oh, say, at least three tries—three being the traditional number, in such matters, equating to 1.6 bits of influence over the future. And even then it might be better to continue despite cries of gerrymandering, rather than give up and let humanity perish. But one should be extremely reluctant to flip through sixteen solutions looking for a good one. If the first try fails, it means that you demanded too little of yourself, failed to rise to the challenge, and you need to rethink your whole strategy before trying again. More than three wishes, and I don't care if you save the human species, you're still incompetent. Such is the advice I would give to the al-Qaeda programmers, along with stern admonishments against checking to make sure humanity followed the way of Islam.

I do think there's light in us *somewhere*, a will that might shape a Nice Place to Live. The question is whether coherent extrapolated volition as an initial dynamic satisfactorily seeks that attractor.

5. Dire Warnings

Don't try this at home.

This says what I currently want. It doesn't say how to do it. The technical side of Friendly AI is not discussed here. The technical side of Friendly AI is hard and requires, like, actual math and stuff.

Don't get caught up in fantasies of the New World Order.

This new version of Friendly AI has an unfortunate disadvantage, which is that it is less vague, and people can speculate about what our extrapolated volitions will want, or argue about it. It will be great fun, and useless, and distracting. Arguing about morality is so much fun that most people prefer it to actually accomplishing anything. This is the same failure that chews up the would-be SI designers with Four Great Moral Principles. If you argue about how your Four Great Moral Principles will be produced by extrapolated volition, it's much the same way to switch off your brain. If you're trying to learn Friendly AI (see HowToLearnFriendlyAI) then you should concentrate on the Friendliness dynamics, and on learning the science background for the technical side. Look to the *structure*, not the *content*, and resist the temptation to argue things that are great fun.

Don't hold the victory party before the battle is won.

Anyone who wants to argue about whether extrapolated volition will favor Democrats or Republicans should recall that *currently* the Earth is scheduled to vanish in a puff of tiny smiley faces, with an unknown deadline and Moore's Law ticking.

As an experiment, I am instituting the following policy on the SL4 mailing list:

None may argue on the SL4 mailing list about the output of CEV, or what kind of world it will create, unless they donate to the Singularity Institute:

- $10 to argue for 48 hours.
- $50 to argue for one month.
- $200 to argue for one year.
- $1000 to get a free pass until the Singularity.

Past donations count toward this total. It's okay to have fun, and speculate, so long as you're not doing it at the expense of actually *helping*.

It goes without saying that anyone wishing to point out a *problem* is welcome to do so. Likewise for talking about the technical side of Friendly AI. It is just the speculation about post-Singularity outcomes and CEV outputs that I am pricing, fearing that otherwise people's souls will be sucked away. (Try to sneak in your fun speculation by adding "which could be a problem if . . . ," and I'll spot it and send you the bill.)

6. Previously Asked Questions about Coherent Extrapolated Volition

Q1. *How do you "extrapolate" what we would think if we knew more, thought faster/smarter? Are we taking the volition of current and historical humans as a data set, and then looking for trends that can be extended outward?* (ChristianRovner)

I was thinking in terms of a detailed model of each individual mind, in as much detail as necessary to guess the class of volition-extrapolating transformations defined by "knew more," "thought faster," etc. In principle, the result we are *approximating* is an idealized version of what would happen if we reached in and performed a set of volition-extrapolating upgrades on your mind: inserted new knowledge, had you sit down and think for a year, watched to see what your opinion would be after growing up for fifty years. But it may be that most of the individual fine details predictably just add to *spread*, or that fine details predictably don't end up *coherent* enough with the rest of humanity to be important to extrapolating the *coherent* extrapolated volition. The key is to set up the underlying question such that approximating the answer is a question of simple fact, to which simple Bayesian intelligence suffices.

Q2. *Removing the ability of humanity to do itself in and giving it a much better chance of surviving Singularity is of course a wonderful goal. But even if you call the FAI "optimizing processes" or some such it will still be a solution outside of humanity rather than humanity growing into being enough to take care of its problems. Whether the FAI is a "parent" or not it will be an alien "gift" to fix what humanity cannot. Why not have humanity itself recursively self-improve?* (SamanthaAtkins)

For myself, the best solution I can imagine at this time is to make CEV our Nice Place to Live, not forever, but to give humanity a breathing space to grow up. Perhaps there is a better way, but this one still seems pretty good. As for it being a solution outside of humanity, or humanity being unable to fix its own problems . . . on this one occasion I say, go ahead and assign the moral responsibility for the fix to the Singularity Institute and its donors.

Moral responsibility for *specific* choices within a CEV is hard to track down, in the era before direct voting. No individual human may have formulated such an intention and acted with intent to carry it out. But as for the *general* fact that a bunch of stuff gets fixed: the programming team and SIAI's donors are human and it was their intention that a bunch of stuff get fixed. I should call this a case of humanity solving its own problems, if on a highly abstract level.

Q3. *Why are you doing this? Is it because your moral philosophy says that what you want is what everyone else wants?* (XR7)

Where would be the base-case recursion? But in any case, no. I'm an individual, and I have my own moral philosophy, which may or may not pay any attention to what our extrapolated volition thinks of the subject. Implementing CEV is just my attempt not to be a jerk.

I do value highly other people getting what they want, among many other values that I hold. But there are certain things such that if people want them, even want them with coherent volitions, I would decline to help; and I think it proper for a CEV to say the same. That is only one person's opinion, however.

Q4. *There are a couple of conclusions that make me very uncomfortable, but unfortunately I can't specify why: "If I later find I'm one of the 5% of humanity whose personal philosophies predictably work out to pure libertarianism, and I threw away my one chance to free humanity—the hell with it." and "Maybe, if it's 80% of humanity in one coherent unit, and 20% in the other coherent unit, the 80% would vote to disenfranchise the 20%, or ensure that the 20% also grew into selfish bastards. As I currently construe CEV, this is a real possibility."* (ChristianRovner)

Worried that CEV isn't doing the right thing?

Yeah. Worried that the possibility is not ruled out by the dynamic.

You're experiencing discomfort because your morality says that certain things are wrong *regardless of what anyone thinks of them*, and here CEV is making it into a majority vote rather than an Inalienable Right. But how does a Friendly Really Powerful Optimization Process know what constitutes "an Inalienable Right, regardless of what anyone thinks of it"? The FAI has to ask *you*—you're the one with this perception of Inalienable Rightness; it isn't recorded anywhere else an FAI can find it. So if you imagine a scenario where a majority of extrapolated humanity doesn't share this perception of an Inalienable Right, then the extrapolated majority might win. I see no way to resolve this discomfort that does not constitute taking over the world, as the al-Qaeda terrorists would sharply advise me if I began writing in a Bill of Rights. The most one can do is appoint a Last Judge with an off switch; predetermining the outcome makes it a farce.

"The FAI has to ask you—you're the one with this perception of Inalienable Rightness; it isn't recorded anywhere else an FAI can find it. So if you imagine a scenario where a majority of humanity doesn't share this perception of an Inalienable Right, then the

majority might win." Are there any inferential steps in between? I don't see it as an obvious sequitur.

Do you want a small minority perception of an Inalienable Right to have power over you? If you don't share it?

Not over me, but over themselves.

That's a structurally determined individual dynamic.

Yes, I guess I'm a pure libertarian. Well. Now I think what worries me is not that CEV is not doing the "right thing." I don't even know if there is a "right thing" to begin with. What worries me is that CEV can exert such kind of coercion over certain individuals. It doesn't sound right.

All I can say is: Last Judge, or whatever FAI theory dictates for verification. CEV has the theoretical power to do arbitrarily awful things if the extrapolation dynamic *doesn't* seek out the light in the heart of humanity, but the Last Judge has to say "Yes" first. And again, I see no way around this without taking over the world. Either humanity has the power of determination, or the programmers do. You can refuse to participate, but not exert detailed control over the outcome.

I don't impose my libertarianism on others. Libertarianism is about not imposing your volition on others.

Right! Now, you just need to take that one step further, and say, "I won't impose my libertarianism on *humankind*." Like I said, there's a fundamental moral question here. And if libertarianism is correct, *I am in fact wrong* to design an initial dynamic that allows for coercion. The reason for the collective dynamic is that you can go collective → individual, but not the other way 'round. If you could go individual → collective but not collective → individual, I'd advocate an individual dynamic. It's all about moral caution. If libertarianism is correct, I am *still* wrong, even taking that argument into account. But I'm not sure. Hence, the moral caution.

I used to be a "protected memory" libertarian, ruling out even the smallest nonconsensual touch. I learned more, thought longer, and now I'm not sure. Nonconsensual touches aren't *that* unbearable a possibility. Eliminate nonconsensual effect and a hell of a lot of human nature goes kerplooey. And it isn't all inhumane nature, either, as I judge humaneness. If there's one piece of advice I have for humanity after the Singularity, it would be, "Take it *sloooow*." I'm a lot more careful with my Humanity points than I used to be. Call me a neanderthal conservative.

What do you mean by "imposing libertarianism"? Rewiring their mind so that they want to be libertarian? Forcing them to live in a society where they can't impose their volition on others? #2, I guess.

Making the choice for libertarianism for humanity.

Jeez. I think I understand. But it still feels uncomfortable.

No one promised that moral caution would be comfortable. Anyone can not take over the world when it seems like a bad idea. It's not taking over the world when it seems like a *good* idea that requires courage. A simple rule for FAI programmers is to never commit any act without giving some humane higher intelligence a veto. Thou art not that smart.

My guess is that the *successor* to the initial dynamic will have a built-in Bill of Rights, ruling out deadly darknesses. Maybe the Bill of Rights will impose that invariant on any successor system humankind votes for, *forever*, even if we change our minds later. The will of humankind that shapes the dynamic in its first moments may dare to exercise that command over the unimaginably far future, for that we would sooner destroy ourselves, revoke the gift we give to our descendants, than let such things come to pass.

But it will not be in the initial dynamic, for I am not that wise.

For *every one* of those Inalienable Rights, the majority will say afterward that the initial dynamic should never have been written to make those Rights subject to majority vote, for Inalienable Rights are Inalienable Rights regardless of what anyone thinks of them. And if I myself dared to set down ten Rights, I would get at least three Wrong.

Q5. *What if it's not possible to model people without simulating them in such fine detail that the simulation becomes a person?* (Frequently Asked)

I can imagine and extrapolate people, at least approximately, without creating new people in my imagination. People cannot be simulated perfectly, regardless of computing power or ethical constraints. There are quantum divergences, unless we become uploads running on hardware insulated from all measurement of the outside world. Even then there would be no way to perfectly extrapolate humanity in advance, before running the process. For this reason do I speak of *predictable* extrapolations, abstract properties of the outcome that a Really Powerful Optimization Process can predict with reasonably high confidence. Likewise I speak of figuring it both ways and computing the *spread*, and leaving our options open where our extrapolated volitions do not converge.

Requiring that the simulation not be sentient may place an upper bound on predictive accuracy for individuals, but *humankind's* coherent extrapolated volition might still be tractable. It might be possible to predict some *coherent* volitions of a *planetary* population with much greater statistical reliability than would apply to a single individual, and perhaps that would suffice for emergency first aid. I would guess that the more obvious the necessity, the greater the probability that our medium-distance volitions will converge. An approximate extrapolation of humankind's volition might state a 99% confidence that a perfect extrapolation of humankind's volition would show between 78% and 82% of the population agreeing on a certain proposition. That might suffice unto writing the successor dynamic and creating a Nice Place to Live, until direct voting becomes possible.

Or maybe a Really Powerful Optimization Process can predict that five years hence you'll walk into a barbershop in Spokane. It seems unlikely—impossible, actually—but maybe I didn't think of the obvious creative shortcut.

Perfect modeling is not required, only enough to satisfice. But yes, sufficient *spread* could increase the extrapolation's *chaos* to the point that it kills CEV as a solution; it simply would not be possible to speak of what humanity "wanted", even in the way of emergency first aid.

A worrying possibility is that requiring non-sentient simulations might create systematic biases; prevent the extrapolation from getting started; or block extrapolation of our probable decisions, or even possible decisions, on specific moral questions such as "What constitutes a 'sentient' simulation?" Ask me again about this problem in a year.

Q6. *What if someone disagrees with the CEV?* (PhilipSutton)

Speculating about outcomes? I counter-speculate that complete silence would not be an inappropriate response. Imagine the silliness of arguing with your own extrapolated volition. It's not only silly, it's dangerous and harmful; you're setting yourself in opposition to the place you would have otherwise gone, rationalizing counterarguments against a good answer.

If humankind's CEV does something I disagree with, maybe there are reasons I haven't realized, maybe my own extrapolated volition happens to disagree with humanity's on this point, maybe I'm an ornery guy who would find fault no matter what the CEV did, maybe the extrapolation of humankind will turn out to be wrong, maybe a different initial dynamic would have worked out differently. A CEV has no moral authority, unless one is foolish enough to regard it as a god. CEV would just be a thing that a human conceived, humans implemented, and that sorta reflects humanity in an odd but useful way. I am not under the slightest obligation to change my opinions in the direction of the CEV, if it appears that we disagree. I need to make my own decisions, so that someday I can grow up and vote.

Humanity needs emergency first aid, but my guess is that for a CEV to argue with humanity, or even display its internal reasoning process, would hurt us. We should invent our own philosophies, not ask our extrapolated volitions, or someday we'll ask our volitions and get back "The only answering procedure you know is asking your volition." Imaginary gods argue with humans and give them orders and demand allegiance; what we need is a real CEV that will shut up and help.

Q7. *Where is the CEV getting this kind of power?* (Frequently Asked)

A Really Powerful Optimization Process is not a god. A Really Powerful Optimization Process could tear apart a god like tinfoil. Hence the extreme caution.

- What is the Singularity?
- Levels of Organization in General Intelligence

Q8. *A problem is whether the most desirable volition is part of the collective or relatively rare. A rare individual or group of individuals' vision may be a much better goal and may perhaps even be what most humans eventually have as their volition when they get wise enough, smart enough and so on.* (SamanthaAtkins)

"Wise enough, smart enough"—this is just what I'm trying to describe specifically how to extrapolate! "What most humans eventually have as their *decision* when they get wise enough, smart enough, and so on" is *exactly* what their extrapolated *volition* is supposed to guess.

It is you who speaks these words, "most desirable volition," "much better goal." How does the dynamic know what is the "most desirable volition" or what is a "much better goal," if not by looking at you to find the initial direction of extrapolation?

How does the dynamic know what is "wiser," if not by looking at your judgment of wisdom? The order of evaluation cannot be recursive, although the order of evaluation can iterate into a steady state. You cannot ask what definition of flongy you would choose if you were flongier. You can ask what definition of wisdom you would choose if you knew more, thought faster. And then you can ask what definition of wisdom-2 your wiser-1 self would choose on the next iteration (keeping track of the increasing distance). But the extrapolation must climb the mountain of your volition from the foothills of your current self; your volition cannot reach down like a skyhook and lift up the extrapolation.

It is a widespread human perception that some individuals are wiser and kinder than others. Suppose our coherent extrapolated volition does decide to weight volitions by wisdom and kindness—a suggestion I strongly dislike, for it smacks of disenfranchisement. It would still take a majority vote of extrapolated volitions for the initial dynamic to decide how to judge "wisdom" and "kindness." I don't think it wise to tell the initial dynamic to look to whichever humans judge *themselves* as wiser and kinder. And if the programmers define their own criteria of "wisdom" and "kindness" into a dynamic's search for leaders, that is taking over the world by proxy. (You wouldn't want the al-Qaeda programmers doing that, right? Though it *might* work out all right in the end, so long as the terrorists told the initial dynamic to *extrapolate* their selected wise men.)

If we know that we are not the best, then let us extrapolate our volitions in the direction of becoming more the people we wish we were, not concentrate Earth's destiny into the hands of our "best." What if our best are not good enough? We should need to extrapolate them farther. If so, why not extrapolate everyone?

Q9. *How does the dynamic force individual volitions to cohere?* (Frequently Asked)

The dynamic doesn't *force* anything. The engineering goal is to ask what humankind "wants," or rather what we would decide if we knew more, thought faster, were more the people we wished we were, had grown up farther together, etc. "There is nothing which humanity can be said to 'want' in this sense" is a possible answer to this question. Meaning, you took your best shot at asking what humanity wanted, and humanity didn't want anything coherent. But you cannot *force* the extrapolated volitions to cohere by computing some other question than "What does humanity want?" That is fighting alligators instead of draining the swamp—solving an engineering subproblem at the expense of what you meant the project to accomplish.

There are nonobvious reasons our volitions would cohere. In the Middle East the Israelis hate the Palestinians and the Palestinians hate the Israelis; in Belfast the Catholics hate the Protestants and the Protestants hate the Catholics; India hates Pakistan and Pakistan hates India. But Gandhi loves everyone, and Martin Luther King loves everyone, and so their wishes add up instead of cancelling out, coherent like the photons in a laser. One might say that love obeys Bose-Einstein statistics while hatred obeys Fermi-Dirac statistics.

Similarly, disagreements may be predicated on:

- Different guesses to simple questions of fact. ("Knew more" increases coherence.)
- Poor solutions to cognitive problems. ("Think faster" increases coherence.)
- Impulses that would play lesser relative roles in the people we wished we were.
- Decisions we would not want our extrapolated volitions to include.

Suppose that's not enough to produce coherence? Extrapolating humankind's volition, as a moral solution, doesn't require some exact particular set of rules for the initial dynamic. You can't take leave of asking what humanity "wants," but it's all right, albeit dangerous, to try more than one plausible definition of "want." I don't think it's gerrymandering to probe the space of "dynamics that plausibly ask what humanity wants" to find a dynamic that produces a *coherent* output, provided:

- The meta-dynamic looks *only* for coherence, no other constraints.
- The meta-dynamic searches a *small* search space.
- The meta-dynamic satisfices rather than maximizes.
- The dynamic itself doesn't force coherence where none exists.
- A Last Judge peeks at the actual answer and checks it makes sense.

I would not object to an initial dynamic that contained a meta-rule along the lines of: "Extrapolate far enough that our medium-distance wishes cohere with each other and don't interfere with long-distance vetoes. If that doesn't work, try this different order of evaluation. If that doesn't work then fail, because it looks like humankind doesn't want anything."

Note that this meta-dynamic tests one quantitative variable (how far to extrapolate) and one binary variable (two possible orders of evaluation), and *not*, say, all possible orders of evaluation.

Forcing a confident answer is a *huge* no-no in FAI theory. If an FAI discovers that an answering procedure is inadequate, *you do not force it to produce an answer anyway.*

Q10. *How accurate and detailed does the extrapolation of humankind's volition need to be in order to work?* (Frequently Asked)

Some important points that I don't think I emphasized clearly enough in the body of the text:

- The optimization process is trying to extrapolate a *population* volition, not the volition of any one individual. I am visualizing and describing this as an actual, detailed, attempt at extrapolating individual minds and their social interaction; but that is because I am defining what I wish to *approximate*.
- The optimization process is extrapolating a spread-out range of possibilities, not *the* answer. Given the uncertainty of underlying quantum processes, no definite answer may exist even in principle.
- Do we want our coherent extrapolated volition to satisfice, or maximize? My guess is that we want our coherent extrapolated volition to satisfice—to apply emergency first aid to human civilization, but not do humanity's work on our behalf, or decide our futures for us. If so, rather than trying to guess the *optimal* decision of a *specific individual*, the CEV would pick a solution that *satisficed* the *spread of possibilities* for the extrapolated *statistical aggregate* of humankind.

This is another reason not to stand in awe of the judgments of a CEV—a solution that satisfices an extrapolated spread of possibilities for the statistical aggregate of humankind may not correspond to the *best* decision of any *individual*, or even the best vote of any real, actual adult humankind.

If all we want out of CEV is a limited set of emergency assistances, measures on which our superposed possible volitions are predictably mostly coherent, then the extrapolation doesn't *need* to be extremely detailed. The extrapolation might be extremely detailed *anyway*—why not?—but it wouldn't *need* to be detailed. Who knows, maybe the guessing of modern-day individuals, based on extremely gross modeling from a tiny information base, would be enough to correctly predict some aspects!

Perhaps the extrapolation will be fine enough, and detailed enough, that there recognizably emerges something like individual volitions voting. Perhaps not. I think I would enjoy knowing that the extrapolation took into account my details and choices as an individual—though not in any understandable way that would tempt me to manipulate my extrapolated vote to childishly chosen ends. I would still like to know that my personal choices played a role in the extrapolation of humankind's volition. It would lend a sense of meaning and participation. But my guess is that this is not *necessary* for CEV to work.