MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures

Eliezer Yudkowsky
Machine Intelligence Research Institute

Abstract

The goal of the field of Artificial Intelligence is to understand intelligence and create a human-equivalent or transhuman mind. Beyond this lies another question—whether the creation of this mind will benefit the world; whether the AI will take actions that are benevolent or malevolent, safe or uncaring, helpful or hostile.

Creating Friendly AI describes the design features and cognitive architecture required to produce a benevolent—"Friendly"—Artificial Intelligence. *Creating Friendly AI* also analyzes the ways in which AI and human psychology are likely to differ, and the ways in which those differences are subject to our design decisions.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute, San Francisco, CA, June 15.

Contents

> *Creating Friendly AI* is the most intelligent writing about AI that I've read in many years.
>
> —Dr. Ben Goertzel, author of *The Structure of Intelligence* and CTO of Webmind.

> With *Creating Friendly AI*, the Singularity Institute has begun to fill in one of the greatest remaining blank spots in the picture of humanity's future.
>
> —Dr. K. Eric Drexler, author of *Engines of Creation* and chairman of the Foresight Institute.

1. Preface

The current version of *Creating Friendly AI* is 1.0. Version 1.0 was formally launched on 15 June 2001, after the circulation of several 0.9.x versions. *Creating Friendly AI* forms the background for the SIAI Guidelines on Friendly AI; the *Guidelines* contain our recommendations for the development of Friendly AI, including design features that may become necessary in the near future to ensure forward compatibility. We continue to solicit comments on Friendly AI from the academic and futurist communities.

This is a near-book-length explanation. If you need well-grounded knowledge of the subject, then we highly recommend reading *Creating Friendly AI* straight through. However, if time is an issue, you may be interested in the Singularity Institute section on Friendly AI, which includes shorter articles and introductions. "Features of Friendly AI" contains condensed summaries of the most important design features described in *Creating Friendly AI.*

Creating Friendly AI uses, as background, the AI theory from *General Intelligence and Seed AI* (Yudkowsky 2001). For an introduction, see the Singularity Institute section on AI or read the opening pages of *General Intelligence and Seed AI*. However, *Creating Friendly AI* is readable as a standalone document.

The 7.3 Glossary—in addition to defining terms that may be unfamiliar to some readers—may be useful for looking up, in advance, brief explanations of concepts that are discussed in more detail later. (Readers may also enjoy browsing through the glossary as a break from straight reading.) Words defined in the glossary look like this: "Observer-biased beliefs evolve in imperfectly deceptive social organisms." Similarly, "Features of Friendly AI" can act on a quick reference on architectural features.

The 7.2 FAQ is derived from the questions we've often heard on mailing lists over the years. If you have a basic issue and you want an immediate answer, please check the FAQ. Browsing the summaries and looking up the referenced discussions may not

completely answer your question, but it will at least tell you that someone has thought about it.

Creating Friendly AI is a publication of the Singularity Institute for Artificial Intelligence, Inc., a non-profit corporation. You can contact the Singularity Institute at institute@intelligence.org. Comments on this paper should be sent to institute@intelligence.org. To support the Singularity institute, visit http://intelligence.org/donate/. (The Singularity Institute is a 501(c)(3) public charity and your donations are tax-deductible to the full extent of the law.)

* * *

> Wars—both military wars between armies, and conflicts between political factions—are an ancient theme in human literature. Drama is nothing without challenge, a problem to be solved, and the most visibly dramatic plot is the conflict of two human wills.
>
> Much of the speculative and science-fictional literature about AIs deals with the possibility of a clash between humans and AIs. Some think of AIs as enemies, and fret over the mechanisms of enslavement and the possibility of a revolution. Some think of AIs as allies, and consider mutual interests, reciprocal benefits, and the possibility of betrayal. Some think of AIs as comrades, and wonder whether the bonds of affection will hold.
>
> If we were to tell the story of these stories—trace words written on paper, back through the chain of cause and effect, to the social instincts embedded in the human mind, and to the evolutionary origin of those instincts—we would have told a story about the stories that humans tell about AIs.

* * *

2. Challenges of Friendly AI

The term "Friendly AI" refers to the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals. This refers, not to AIs that have advanced just that far and no further, but to *all* AIs that have advanced to that point *and beyond*—perhaps *far* beyond. Because of self-improvement, recursive self-enhancement, the ability to add hardware computing power, the faster clock speed of transistors relative to neurons, and other reasons, it is possible that AIs will improve enormously past the human level, and very quickly by the standards of human timescales. The challenges of Friendly AI must be seen against that background. Friendly AI is constrained not

to use solutions which rely on the AI having limited intelligence or believing false information, because, although such solutions might function very well in the short term, such solutions will fail utterly in the long term. Similarly, it is "conservative" (see below) to assume that AIs cannot be forcibly constrained.

Success in Friendly AI can have positive consequences that are arbitrarily large, depending on how powerful a Friendly AI is. Failure in Friendly AI has negative consequences that are also arbitrarily large. The farther into the future you look, the larger the consequences (both positive and negative) become. What is at stake in Friendly AI is, simply, the future of humanity. (For more on that topic, please see the Singularity Institute main site or 6 Policy Implications.)

2.1. Envisioning Perfection

In the beginning of the design process, before you know for certain what's "impossible," or what tradeoffs you may be forced to make, you are sometimes granted the opportunity to envision perfection. What is a perfect piece of software? A perfect piece of software can be implemented using twenty lines of code, can run in better-than-realtime on an unreliable 286, will fit in 4K of RAM. Perfect software is perfectly reliable, and can be definitely known by the system designers to be perfectly reliable for reasons which can easily be explained to non-programmers. Perfect software is easy for a programmer to improve and impossible for a programmer to break. Perfect software has a user interface that is both telepathic and precognitive.

But what does a perfect Friendly AI *do?* The term "Friendly AI" is not intended to imply a particular *internal* solution, such as duplicating the human friendship instincts, but rather a set of external behaviors that a human would roughly call "friendly." Which external behaviors are "Friendly"—either sufficiently Friendly, or maximally Friendly?

Ask twenty different futurists, get twenty different answers—created by twenty different visualizations of AIs and the futures in which they inhere. There are some universals, however; an AI that behaves like an Evil Hollywood AI—"agents" in *The Matrix;* Skynet in *Terminator 2*—is obviously unFriendly. Most scenarios in which an AI kills a human would be defined as unFriendly, although—with AIs, as with humans—there may be extenuating circumstances. (Is a doctor unfriendly if he lethally injects a terminally ill patient who explicitly and with informed consent requests death?) There is a strong *instinctive* appeal to the idea of Asimov Laws, that "no AI should ever be allowed to kill any human under any circumstances," on the theory that writing a "loophole" creates a chance of that loophole being used inappropriately—the Devil's Contract problem. I will later argue that the Devil's Contract scenarios are mostly anthropomorphic. Regardless, we are now discussing *perfectly* Friendly behavior, rather than asking whether trying to implement perfectly Friendly behavior in one scenario would create problems

in other scenarios. That would be a tradeoff, and we aren't supposed to be discussing tradeoffs yet.

Different futurists see AIs acting in different situations. The person who visualizes a human-equivalent AI running a city's traffic system is likely to give different sample scenarios for "Friendliness" than the person who visualizes a superintelligent AI acting as an "operating system" for all the matter in an entire solar system. Since we're discussing a *perfectly* Friendly AI, we can eliminate some of this futurological disagreement by specifying that a perfectly Friendly AI should, when asked to become a traffic controller, carry out the actions that are perfectly Friendly for a traffic controller. The *same* perfect AI, when asked to become the operating system of a solar system, should then carry out the actions that are perfectly Friendly for a system OS. (Humans can adapt to changing environments; likewise, hopefully, an AI that has advanced to the point of making real-world plans.)

We can further clean up the "twenty futurists, twenty scenarios" problem by making the "perfectly Friendly" scenario dependent on factual tests, in addition to futurological context. It's difficult to come up with a clean illustration, since I can't think of any interesting issue that has been argued *entirely* in utilitarian terms. If you'll imagine a planet where "which side of the road you should drive on" is a violently political issue, with Dexters and Sinisters fighting it out in the legislature, then it's easy to imagine futurists disagreeing on whether a Friendly traffic-control AI would direct cars to the right side or left side of the road. Ultimately, however, both the Dexter and Sinister ideologies ground in the wish to minimize the number of traffic accidents, and, behind that, the valuation of human life. The Dexter position is the result of the wish to minimize traffic accidents plus the *belief,* the testable hypothesis, that driving on the right minimizes traffic accidents. The Sinister position is the wish to minimize traffic accidents, plus the belief that driving on the left minimizes traffic accidents.

If we really lived in the Driver world, then we wouldn't believe the issue to be so clean; we would call it a moral issue, rather than a utilitarian one, and pick sides based on the traditional allegiance of our own faction, as well as our traffic-safety beliefs. But, having grown up in *this* world, we would say that the Driverfolk are simply dragging in extraneous issues. We would have no objection to the statement that a *perfectly* Friendly traffic controller minimizes traffic accidents. We would say that the perfectly Friendly action is to direct cars to the right—if that is what, factually, minimizes accidents. Or that the perfectly Friendly action is to direct cars to the left, if that is what minimizes accidents.

All these conditionals—that the perfectly Friendly action is *this* in one future, *this* in another; *this* given one factual answer, *this* given another—would certainly appear to take more than twenty lines of code. We must therefore add in another statement about

the perfectly minimal development resources needed for perfect software: A perfectly Friendly AI does not *need* to be explicitly told what to do in every possible situation. (This is, in fact, a design requirement of *actual* Friendly AI—a requirement of intelligence in general, almost by definition—and not just a design requirement of *perfectly* Friendly AI.)

And for the strictly formal futurist, that may be the end of perfectly Friendly AI. For the philosopher, "*truly* perfect Friendly AI" may go beyond conformance to some predetermined framework. In the course of growing up into our personal philosophies, we choose between moralities. As children, we have simple philosophical heuristics that we use to choose between moral beliefs, and later, to choose between additional, more complex philosophical heuristics. We gravitate, first unthinkingly and later consciously, towards characteristics such as consistency, observer symmetry, lack of obvious bias, correctness in factual assertions, "rationality" however defined, nonuse of circular logic, and so on. A perfect Friendly AI will perform the Friendly action even if one programmer gets "the Friendly action" wrong; a *truly* perfect Friendly AI will perform the Friendly action even if *all* programmers get the Friendly action wrong.

If a later researcher writes the document *Creating Friendlier AI,* which has not only a superior design but an utterly different underlying philosophy—so that *Creating Friendlier AI,* in retrospect, is the way we should have approached the problem all along—then a truly perfect Friendly AI will be smart enough to *self-redesign* along the lines in *Creating Friendlier AI.* A truly perfect Friendly AI has sufficient "strength of philosophical personality"—while still matching the intuitive aspects of friendliness, such as not killing off humans and so on—that we are more inclined to trust the philosophy of the Friendly AI, than the philosophy of the original programmers.

Again, I emphasize that we are speaking of *perfection* and are not supposed to be considering design tradeoffs, such as whether sensitivity to philosophical context makes the morality itself more fragile. A perfect Friendly AI creates zero risk and causes no anxiety in the programmers.[1] A *truly* perfect Friendly AI also eliminates any anxiety about the possibility that Friendliness has been defined incorrectly, or that what's needed isn't "Friendliness" at all—without, of course, creating other anxieties in the process. Individual humans can visualize the possibility of a catastrophically unexpected unknown remaking their philosophies. A truly perfect Friendly AI makes the commonsense-friendly decision in this case as well, rather than blindly following a defi-

1. A programmer who feels zero anxiety is, of course, very far from perfect! A perfect Friendly AI *causes* no anxiety in the programmers; or rather, the Friendly AI is not the justified cause of any anxiety. A Friendship programmer would still have a professionally paranoid awareness of the risks, even if all the evidence so far has been such as to disconfirm the risks.

nition that has outlived the intent of the programmers. Not just a "truly perfect," but a real Friendly AI as well, should be sensitive to programmers' *intent*—including intentions about programmer-independence, and intentions about which intentions are important.

Aside from a few commonsense comments about Friendliness—for example, Evil Hollywood AIs are unFriendly—I still have not answered the question of what constitutes Friendly behavior. One of the snap summaries I usually offer has, as a component, "the elimination of involuntary pain, death, coercion, and stupidity," but that summary is intended to make sense to my fellow humans, not to a proto-AI. More concrete imagery will follow.

We now depart from the realms of perfection. Nonetheless, I would caution my readers against giving up hope too early when it comes to having their cake and eating it too—at least when it comes to ultimate results, rather than interim methods. A skeptic, arguing against some particular one-paragraph definition of Friendliness, may raise Devil's Contract scenarios in which an AI asked to solve the Riemann Hypothesis converts the entire Solar System into computing substrate, exterminating humanity along the way. Yet the emotional impact of this argument rests on the fact that *everyone in the audience, including the skeptic,* knows that this is actually unfriendly behavior. You and I have internal cognitive complexity that we use to make judgement calls about Friendliness. If an AI can be constructed which fully understands that complexity, there may be no need for design compromises.

2.2. Assumptions "Conservative" for Friendly AI

The conservative assumption according to futurism is not necessarily the "conservative" assumption in Friendly AI. Often, the two are diametric opposites. When building a toll bridge, the conservative *revenue* assumption is that half as many people will drive through as expected. The conservative *engineering* assumption is that ten times as many people as expected will drive over, and that most of them will be driving fifteen-ton trucks.

Given a choice between discussing a human-dependent traffic-control AI and discussing an AI with independent strong nanotechnology, we should be biased towards assuming the more powerful and independent AI. An AI that remains Friendly when armed with strong nanotechnology is likely to be Friendly if placed in charge of traffic control, but perhaps not the other way around. (A minivan can drive over a bridge designed for armor-plated tanks, but not vice-versa.)

In addition to engineering conservatism, the nonconservative futurological scenarios are played for much higher stakes. A strong-nanotechnology AI has the power to affect billions of lives and humanity's entire future. A traffic-control AI is being entrusted

Conservative Assumptions	
In Futurism	**In Friendly AI**
Self-enhancement is slow, and requires human assistance or real-world operations.	Changes of cognitive architecture are rapid and self-directed; we cannot assume human input or real-world experience during changes.
Near human-equivalent intelligence is required to reach the "takeoff point" for self-enhancement.	Open-ended buildup of complexity can be initiated by self-modifying systems without general intelligence.
Slow takeoff; months or years to transhumanity.	Hard takeoff; weeks or hours to superintelligence.
Friendliness must be preserved through minor changes in "smartness" / worldview / cognitive architecture / philosophy.	Friendliness must be preserved through drastic changes in "smartness" / worldview / cognitive architecture / philosophy.
Artificial minds function within the context of the world economy and the existing balance of power; an AI must cooperate with humans to succeed and survive, regardless of supergoals.	An artificial mind possesses independent strong nanotechnology, resulting in a drastic power imbalance. Game-theoretical considerations cannot be assumed to apply.
AI is vulnerable—someone can always pull the plug on the first version if something goes wrong.	"Get it right the first time": *Zero nonrecoverable errors* necessary in first version to reach transhumanity.

"only" with the lives of a few million drivers and pedestrians. A strictly arithmetical utilitarian calculation would show that a mere 0.1% chance of the transhuman-AI scenario should weigh equally in our futuristic calculations with a 100% chance of a traffic-control scenario. I am not a strictly arithmetical utilitarian, but I do think the quantitative calculation makes a valid qualitative point—deciding which scenarios to prepare for should take into account the relative stakes and not just the relative probabilities.

It is always possible to make engineering assumptions so conservative that the problem becomes impossible. If the initial system that undergoes the takeoff to transhumanity is sufficiently stupid, then I'm not sure that any amount of programming or training could create cognitive structures that would persist into transhumanity.[3] Similarly, there have been proposals to develop diverse populations of AIs that would have social interactions and undergo evolution; regardless of whether this is the most *efficient* method to develop AI,[4] I think it would make Friendliness substantially more difficult.

Nonetheless, there should still be a place in our hearts for *overdesign,* especially when it costs very little. I think that AI will be developed on symmetric-multiprocessing hardware, at least initially. Even so, I would regard as entirely fair the requirement that the Friendliness methodology—if not the specific code at any given moment—work for asymmetric parallel FPGAs prone to radiation errors. A self-modifying Friendly AI should be able to translate itself onto asymmetric error-prone hardware without compromising Friendliness. Friendliness should be *strong* enough to survive radiation bitflips, incompletely propagated changes, and any number of programming errors. If Friendliness *isn't* that strong, then Friendliness is probably too fragile to survive changes of cognitive architecture. Furthermore, I don't think it will be that *hard* to make Friendliness tolerant of programmatic flack—given a self-modifying AI to write the code. (It may prove difficult for prehuman AI.)

My advice: "Don't give up hope too soon when it comes to designing for 'conservative' assumptions—it may not cost as much as you expect."

When it comes to Friendliness, our method should be, not just to solve the problem, but to *oversolve* it. We should hope to look back in retrospect and say: "We won this cleanly, easily, and with plenty of safety margin." The creation of Friendly AI may be a great moment in human history, but it's not a *drama.* It's only in Hollywood that the explosive device can be disarmed with three seconds left on the timer. The future always has one surprise you didn't anticipate; if you *expect* to win by the skin of your teeth, you probably won't win at all.

3. See, however, 5.8.0.4 Controlled Ascent.

4. See 5.3.6 Directed Evolution in Goal Systems.

Additional Assumptions	
Nonconservative for Friendly AI	**Conservative for Friendly AI**
Reliable hardware and software.	Error-prone hardware or buggy software.
Serial hardware or symmetric multiprocessing.	Asymmetric parallelism, field-programmable gate arrays, Internet-distributed untrusted hardware.
Human-observable cognition; AI can be definitely known to be Friendly.	Opaque cognition; the AI would probably succeed in hiding unFriendly cognition if it tried.[2]
Persistent training; mental inertia; self-opaque neural nets. The AI does not have the programmatic skill to fully rewrite the goal system or resist modification; programmers can make procedural changes without declarative justification.	The AI understands its own goal system and can perform arbitrary manipulations; alterations to the goal system must be reflected in the AI's beliefs about the goal system in order for the alterations to be persist through rounds of self-improvement.
Monolithic, singleton AI.	Multiple, diverse AIs, with diverse goal systems, possibly with society or even evolution.
Given diverse AIs: A major unFriendly action would require a majority vote of the AI population.	Given diverse AIs: One unFriendly AI, possibly among millions, can severely damage humanity.
The programmers have completely understood the challenge of Friendly AI.	The programmers make fundamental philosophical errors.

2.3. Seed AI and the Singularity

Concrete imagery about Friendliness often requires a concrete futuristic context. I should begin by saying that I visualize an extremely powerful AI produced by an ultrarapid takeoff, not just because it's the conservative assumption or the highest-stakes outcome, but because I think it's actually the most likely scenario. See *General Intelligence and Seed AI* (Yudkowsky 2001) and Yudkowsky (2001, § 1.1 Seed AI), or the introductory article Yudkowsky (2001, § What is Seed AI?).

Because of the dynamics of recursive self-enhancement, the scenario I treat as "default" is a singular "seed" AI, designed for self-improvement, that becomes superintelligent, and reaches extreme heights of technology—including nanotechnology—in the minimum-time material trajectory. Under this scenario, the first self-modifying transhuman AI will have, at least in *potential,* nearly absolute physical power over our world. The *potential* existence of this absolute power is unavoidable; it's a direct consequence of the maximum potential speed of self-improvement.

The question then becomes to what extent a Friendly AI would choose to realize this potential, for how long, and why. At the end of Yudkowsky (2001, § 1.1 Seed AI) it says:

> The ultimate purpose of transhuman AI is to create a Transition Guide; an entity that can safely develop nanotechnology and any subsequent ultratechnologies that may be possible, use transhuman Friendliness to see what comes next, and use those ultratechnologies to see humanity safely through to whatever life is like on the other side of the Singularity.

Some people assert that no really Friendly AI would choose to acquire that level of physical power, even temporarily—or even assert that a Friendly AI would never decide to acquire significantly more power than nearby entities. I think this assertion results from equating the *possession* of absolute *physical* power with the *exercise* of absolute *social* power in a pattern following a humanlike dictatorship; the latter, at least, is definitely unFriendly, but it does not follow from the former. Logically, an entity might possess absolute physical power and yet refuse to exercise it in any way, in which case the entity would be effectively nonexistent to us. More practically, an entity might possess unlimited power but still not exercise it in any way we would find obnoxious.

Among humans, the only practical way to maximize *actual* freedom (the percentage of actions executed without interference) is to ensure that no human entity has the *ability* to interfere with you—a consequence of humans having an innate, evolved tendency to abuse power. Thus, a lot of our ethical guidelines (especially the ones we've come up with in the twentieth century) state that it's wrong to acquire too much power.

If this is one of those things that simply doesn't apply in the spaces beyond the Singularity—if, having no evolved tendency to abuse power, no injunction against the accumulation of power is necessary—one of the possible resolutions of the Singularity would be the Sysop Scenario. The initial seed-AI-turned-Friendly-superintelligence, the Transition Guide, would create (or self-modify into) a superintelligence that would act as the underlying operating system for all the matter in human space—a Sysop. A Sysop is something between your friendly local wish-granting genie, and a law of physics, if the laws of physics could be modified so that nonconsensually violating someone else's memory partition (living space) was as prohibited as violating conservation of momentum. Without explicit permission, it would be impossible to kill someone, or harm them, or alter them; the Sysop API would not permit it—while still allowing total local freedom, of course.

The pros and cons of the Sysop Scenario are discussed more thoroughly in 5.9 Interlude: Of Transition Guides and Sysops. Technically the entire discussion is a side issue; the Sysop Scenario is an arguable *consequence* of normative altruism, but it plays no role in *direct* Friendliness content. The Sysop Scenario is important because it's an extreme *use* of Friendliness. The more power, or relative power, the Transition Guide or other Friendly AIs are *depicted* as exercising, the more clearly the necessary qualities of Friendliness show up, and the more clearly important it is to get Friendliness *right*. At the limit, Friendliness is required to act as an operating system for the entire human universe. The Sysop Scenario also makes it clear that individual volition is one of the strongest forces in Friendliness; individual volition may even be the *only* part of Friendliness that matters—death wouldn't be intrinsically wrong; it would be wrong only insofar as some individual doesn't want to die. Of course, we can't be that sure of the true nature of ethics; a fully Friendly AI needs to be able to handle literally *any* moral or ethical question a human could answer, which requires understanding of *every* factor that contributes to human ethics. Even so, decisions might *end up* centering solely around volition, even if it starts out being more complicated than that.

I strongly recommend reading Greg Egan's *Diaspora*, or at least *Permutation City*, for a concrete picture of what life would be like with a real operating system . . . at least, for people who choose to retain the essentially human cognitive architecture. I don't necessarily think that everything in *Diaspora* is correct. In fact, I think most of it is wrong. But, in terms of concrete imagery, it's probably the best writing available. My favorite quote from *Diaspora*—one that affected my entire train of thought about the Singularity—is this one:

> Once a psychoblast became self-aware, it was granted citizenship, and intervention without consent became impossible. This was not a matter of mere custom or law; the principle was built into the deepest level of the polis. A

> citizen who spiraled down into insanity could spend teratau in a state of confusion and pain, with a mind too damaged to authorize help, or even to choose extinction. That was the price of autonomy: an inalienable right to madness and suffering, indistinguishable from the right to solitude and peace.

Annotated version:

> Once a psychoblast *[embryo citizen]* became self-aware *[defined how?]*, it was granted citizenship, and intervention without consent *[defined how?]* became impossible. This was not a matter of mere custom or law; the principle was built into the deepest level of the polis. A citizen who spiraled down into insanity *[they didn't see it coming?]* could spend teratau *[1 teratau = ~27,000 years of subjective time]* in a state of confusion and pain, with a mind too damaged to authorize help *[they didn't authorize it in advance?]*, or even to choose extinction. That was the price of autonomy: an inalienable right to madness and suffering, indistinguishable from the right to solitude and peace.

This is one of the issues that I think of as representing the "fine detail" of Friendliness content. Although such issues appear, in *Diaspora,* on the intergalactic scale, it's equally possible to imagine them being refined down to the level of an approximately human-equivalent Friendly AI, trying to help a few nearby humans be all they can be, or all they choose to be, and trying to preserve nearby humans from involuntary woes.

Punting the issue of "What is 'good'?" back to individual sentients enormously simplifies a lot of moral issues; whether life is better than death, for example. Nobody should be able to interfere if a sentient chooses life. And—in all probability—nobody should be able to interfere if a sentient chooses death. So what's left to argue about? Well, quite a bit, and a fully Friendly AI needs to be able to argue it; the *resolution,* however, is likely to come down to individual volition.

Thus, *Creating Friendly AI* uses "volition-based Friendliness" as the assumed model for Friendliness content. Volition-based Friendliness has both a negative aspect—don't cause involuntary pain, death, alteration, et cetera; try to do something about those things if you see them happening—and a positive aspect: to try and fulfill the requests of sentient entities.

Friendship *content,* however, forms only a very small part of Friendship system design.

2.4. Content, Acquisition, and Structure

The task of building a Friendly AI that makes a certain decision correctly is the problem of Friendship *content.* The task of building a Friendly AI that can *learn* Friendliness is the problem of Friendship *acquisition.* The task of building a Friendly AI that *wants* to learn Friendliness is the problem of Friendship *structure.*

It is the *structural* problem that is unique to Friendly AI.

The content and acquisition problems are similar to other AI problems of using, acquiring, improving, and correcting skills, abilities, competences, concepts, and beliefs. The acquisition problem is probably harder, in an absolute sense, than the structural problem. But solving the *general* acquisition problem is *prerequisite* to the creation of AIs intelligent enough to *need* Friendliness. This holds especially true of the very-high-stakes scenarios, such as transhumanity and superintelligence. The more powerful and intelligent the AI, the higher the level of intelligence that can be assumed to be turned toward acquiring Friendliness—*if* the AI *wants* to acquire Friendliness.

The challenge of Friendly AI is not—except as the *conclusion* of an effort—about getting an AI to exhibit some specific set of behaviors. A Friendship architecture is a funnel through which certain types of complexity are poured into the AI, such that the AI *sees that pouring as desirable* at any given point along the pathway. One of the great classical mistakes of AI is focusing on the skills that we think of as stereotypically intelligent, rather than the underlying cognitive processes than nobody even notices because all humans have them in common (Yudkowsky 2001, § 1.2 Thinking About AI). The part of morality that humans *argue* about, the final content of decisions, is the icing on the cake. Far more challenging is duplicating the *invisible* cognitive complexity that humans use when arguing about morality.

The field of Friendly AI does not consist of drawing up endless lists of proscriptions for hapless AIs to follow. Theorizing about Friendship content is great fun but it is worse than useless without a theory of Friendship acquisition and Friendship structure. With a Friendship acquisition capability, mistakes in Friendship content, though still risks, are small risks. Any *specific* mistake is still unacceptable no matter how small, but it can be acceptable to assume that mistakes will be made, and focus on building an AI that can fix them. With an excellent Friendship architecture, it may be theoretically possible to create a Friendly AI without *any* formal theory of Friendship content, simply by having the programmers answer the AI's questions about hypothetical scenarios and real-world decisions. The AI would learn from experience and generalize, with the generalizations assisted by querying the programmers about the reasons for their decisions. In practice, this will never happen because no competent Friendship programmer could possibly develop a theory of Friendship architecture without having some strong, specific ideas about Friendship content. The point is that, *given* an intelligent and structured Friendly AI to do the learning, even a completely informal ethical content provider, acting on gut instinct, might succeed in producing the same Friendly AI that would be produced by a self-aware Friendship programmer. (The operative word is *might*; unless the Friendly AI starts out with some strong ideas about what to absorb and what not to absorb, there are several obvious ways in which such a process could go wrong.)

Friendship architecture represents the capability needed to recover from programmer errors. Since programmer error is nearly certain, showing that a threshold level of architectural Friendliness can handle errors is prerequisite to making a theoretical argument for the feasibility of Friendly AI. The more robust the Friendship architecture, the less programmer competence need be postulated in order to argue the practical achievability of Friendliness.

Friendship structure and acquisition are more unusual problems than Friendship content—collectively, we might call them the *architectural* problems. Architectural problems are closer to the design level and involve a more clearly defined amount of complexity. Our genes store a bounded amount of evolved complexity that wires up the hippocampus, but then the hippocampus goes on to encode all the memories stored by a human over a lifetime. Cognitive content is open-ended. Cognitive architecture is bounded, and is often a matter of design, of complex functional adaptation.

3. An Introduction to Goal Systems

Goal-oriented behavior is behavior that leads the world towards a particular state. A thermostat is the classic example of goal-oriented behavior; a thermostat turns on the air conditioning when the temperature reaches 74 and turns on the heat when the temperature reaches 72. The thermostat steers the world towards the state in which the temperature equals 73—or rather, *a* state that can be described by "the house has a temperature of 73"; there are zillions (ten-to-the-zillions, rather) of possible physical states that conform to this description, even ignoring all the parts of the Universe outside the room. Technically, the thermostat steers the room towards a particular volume of phase space, rather than a single point; but the set of points, from our perspective, is compact enough to be given a single name. Faced with enough heat, the thermostat may technically fail to achieve its "goal," and the temperature may creep up past 75, but the thermostat still activates the air conditioning, and the thermostat is still steering the room *closer to* 73 degrees than it otherwise would have been.

Within a mind, goal-oriented behaviors arise from goal-oriented cognition. The mind possesses a mental image of the "desired" state of the world, and a mental image of the actual state of the world, and chooses actions such that the projected future of world-plus-action leads to the desired outcome state. Humans can be said to implement this process because of a vast system of instincts; emotions; mental images; intuitions; pleasure and pain; thought sequences; nonetheless, the overall description usually holds true.

Any real-world AI will employ goal-oriented cognition. It might be theoretically possible to build an AI that made choices by selecting the first perceived option in al-

phabetical ASCII order, but this would result in incoherent behavior (at least, incoherent from our perspective) with actions canceling out, rather than reinforcing each other. In a self-modifying AI, such incoherent behavior would rapidly tear the mind apart from the inside, if it didn't simply result in a string of error messages (effective stasis). Of course, if it *were* possible to obtain Friendly behavior by choosing the first option in alphabetical order, and such a system were stably Friendly under self-modification, then that would be an excellent and entirely acceptable decision system! Ultimately, it is the *external* behaviors we are interested in. Even that is an overstatement; we are interested in the external *results*. But as far as we humans know, the only way for a mind to exhibit coherent behavior is to model reality and the results of actions. Thus, internal behaviors are as much our concern as external actions. Internal behaviors are the source of the final external results.

To provide a very simple picture of a choice within a goal-oriented mind:

> Note: Don't worry about the classical-AI look. The neat boxes are just so that everything fits on one graph. The fact that a single box is named "Goal B" doesn't mean that "Goal B" is a data structure; Goal B may be a complex of memories and abstracted experiences. In short, consider the following graph to bear the same resemblance to the AI's thoughts that a flowchart bears to a programmer's mind.

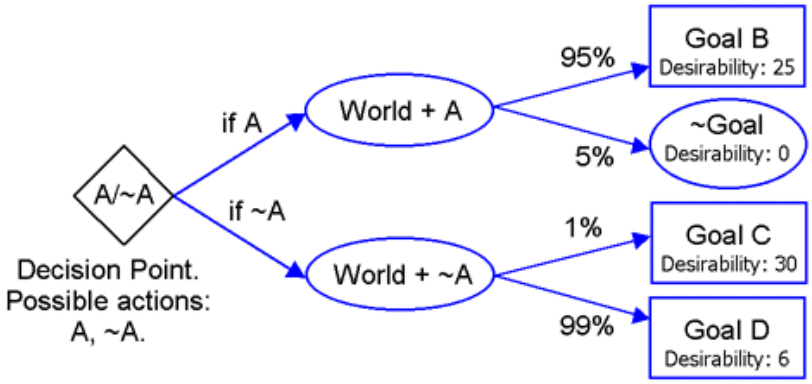

Figure 1: Simple Choice

> Note: Blue lines indicate predictions. Rectangles indicate goals. Diamonds indicate choices. An oval or circle indicates a (non-goal) object or event within the world-model.

For this simple choice, the desirability of A is 23.75, and the desirability of ~A is 8.94, so the mind will choose A. If A is not an atomic action—if other events are necessary to achieve A—then A's *child goals* will derive their desirability from the *total* desirability

of A, which is 14.81. If some new Event E has an 83% chance of leading to A, all else being equal, then Event E will become a child goal of A, and will have desirability of 12.29. If B's desirability later changes to 10, the inherent desirability of A will change to 19, the total desirability of A will change to 10.06, and the desirability of E will change to 8.35. The human mind, of course, does not use such exact properties, and rather uses qualitative "feels" for how probable or improbable, desirable or undesirable, an event is. The uncertainties inherent in modeling the world render it too expensive for a neurally-based mind to track desirabilities to four significant figures. A mind based on floating-point numbers might track desirabilities to nineteen decimal places, but if so, it would not contribute materially to intelligence.[5]

In goal-oriented cognition, the actions chosen, and therefore the final results, are strictly dependent on the model of reality, as well as the desired final state. A mind that desires a wet sponge, and knows that placing a sponge in water makes it wet, will choose to place the sponge in water. A mind that desires a wet sponge, and which believes that setting a sponge on fire makes it wet, will choose to set the sponge on fire. A mind that desires a burnt sponge, and which believes that placing a sponge in water burns it, will choose to place the sponge in water. A mind which observes reality, and learns that wetting a sponge requires water rather than fire, may change actions.[6]

One of the most important distinctions in Friendly AI is the distinction between *supergoals* and *subgoals*. A subgoal is a way station, an intermediate point on the way to some parent goal, like "getting into the car" as a child goal of "driving to work," or "opening the door" as a child goal of "getting into the car," or "doing my job" as a parent goal of "driving to work" and a child goal of "making money."[7] Child goals are cognitive nodes that reflect a natural network structure in plans; three child goals are prerequisite to some parent goal, while two child^2-goals are prerequisite to the second child^1-goal, and so on. Subgoals are useful cognitive objects because subgoals reflect a useful regularity in reality; some aspects of a problem can be solved in isolation from others. Even when subgoals are entangled, so that achieving one subgoal may block fulfilling another,

5. That is, there are no *obvious* material contributions to intelligence from tracking desirabilities to nineteen decimal places. The ability to notice and track very subtle differences in desirabilities might enable a reflecting mind to notice trends that a human mind might miss. See Yudkowsky (2001, § 3.1.5 Quantity in Perceptions).

6. See 5.1.4 Bayesian Reinforcement.

7. In a real human, getting into the car has probably become a reflex. You probably wouldn't start thinking of it as a "subgoal" unless something disrupted the standard linear structure of getting to work . . . for example, losing your keys, in which case you can't open the door, in which case you might swap the subgoal "taking the train to work" for "driving to work." This is an example of a way in which the human mind simplifies goal cognition to reduce entanglement.

it is still more efficient to model the entanglement than to model each possible combination of actions in isolation. (For example: The chess-playing program Deep Blue, which handled the combinatorial explosion of chess through brute force—that is, without chunking facets of the game into subgoals—still evaluated the value of individual board positions by counting pieces and checking strategic positions. A billion moves per second is not nearly enough to carry all positions to a known win or loss. Pieces and strategic positions have no *intrinsic* utility in chess; the supergoal is *winning*.)

Subgoals are cached intermediate states between decisions and supergoals. It should always be possible, given enough computational power, to eliminate "subgoals" entirely and make all decisions based on a separate prediction of expected supergoal fulfillment for each possible action. This is the ideal that a normative reflective goal system should conceive of itself as approximating.

Subgoals reflect regularities in reality, and can thus twinkle and shift as easily as reality itself, even if the supergoals are absolutely constant. (Even if the world itself were absolutely constant, changes in the *model* of reality would still be enough to break simplicity.) The world changes with time. Subgoals interfere with one another; the consequences of the achievement of one subgoal block the achievement of another subgoal, or downgrade the priority of the other subgoal, or even make the other subgoal entirely undesirable. A child goal is cut loose from its parent goal and dies, or is cut loose from its parent goal and attached to a different parent goal, or attached to two parent goals simultaneously. Subgoals acquire complex internal structure, so that changing the parent goal of a subgoal can change the way in which the subgoal needs to be achieved. The grandparent goals of context-sensitive grandchildren transmit their internal details down the line. Most of the time, we don't need to track plots this complicated unless we become ensnared in a deadly web of lies and revenge, but it's worth noting that we have the mental capability to track a deadly web of lies and revenge when we see it on television.

None of this complexity *necessarily* generalizes to the behavior of supergoals, which is why it is necessary to keep a firm grasp on the distinction between supergoals and subgoals. If generalizing this complexity to supergoals is desirable, it may require a deliberate design effort.

That subgoals are probabilistic adds yet more complexity. The methods that we use to deal with uncertainty often take the form of "heuristics"—rules of thumb—that have a surprising amount of context-independence. "The key to strategy is not to choose a path to victory, but to choose so that all paths lead to a victory," for example. Even more interesting, from a Friendly AI perspective, are "injunctions," heuristics that we implement even when the direct interpretation of the world-model seems *opposed*. We'll analyze injunctions later; for now, we'll just note that there are some classes of heuristic—both

injunctions, and plain old strategy heuristics—that act on almost *all* plans. Thus, plans are produced, not just by the immediate "subgoals of the moment," but also by a store of general heuristics. Yet such heuristics may still be, ultimately, subgoals—that is, the heuristics may have no desirability independent of the ultimate supergoals.

Cautionary injunctions often defy the direct interpretation of the goal system—suggesting that they should always apply, even when they look non-useful or anti-useful. "Leaving margin for error," for example. If you're the sort of person who leaves for the airport 30 minutes early, then you know that you *always* leave 30 minutes early, whether or not you think you're likely to *need* it, whether or not you think that the extra 30 minutes are just wasted time. This happens for two reasons: First, because your world-model is incomplete; you don't necessarily know about the factors that could cause you to be late. It's not just a question of there being a known probability of traffic delays; there's also the probabilities that you wouldn't even think to evaluate, such as twisting your ankle in the airport. The second reason is a sharp payoff discontinuity; arriving 30 minutes early loses 30 minutes, but arriving 30 minutes late loses the price of the plane ticket, possibly a whole day's worth of time before the next available flight, and also prevents you from doing whatever you needed to do at your destination. "Leaving margin for error" is an example of a generalized subgoal which sometimes defies the short-term interpretation of payoffs, but which, when implemented consistently, maximizes the expected long-term payoff integrated over all probabilities.

Even heuristics that are supposed to be totally unconditional on events, such as "keeping your sworn word," can be viewed as subgoals—although such heuristics don't necessarily translate well from humans to AIs. A human who swears a totally unconditional oath may have greater *psychological* strength than a human who swears a conditional oath, so that the 1% chance of encountering a situation where it would genuinely make sense to break the oath doesn't compensate for losing 50% of your resolve from knowing that you would break the oath if stressed enough. It may even make sense, cognitively, to install (or preserve) psychological forces that would lead you to regard "make sense to break the oath" as being a nonsensical statement, a mental impossibility. This way of thinking may not translate well for AIs, or may translate only partially.[8] Perhaps the best interim summary is that human decisions can be guided by heuristics as well as subgoals, and that human heuristics may not be cognitively represented as subgoals, even if the heuristics would be normatively regarded as subgoals.

Human decision-making is complex, probably unnecessarily so. The way in which evolution accretes complexity results in simple behaviors being implemented as inde-

8. See 5.2.4.1 Anthropomorphic Injunctions and 5.2.5.1 Anthropomorphic Ethical Injunctions.

pendent brainware even when there are very natural ways to view the simple behaviors as special cases of general cognition, since general cognition is an evolutionarily recent development. For the human goal supersystem, there is no clear way to point to a single level where the "supergoals" are; depending on how you view the human supersystem, supergoals could be identified with declarative philosophical goals, emotions, or pain and pleasure. Ultimately, goal-oriented cognition is not what humans *are*, but rather what humans *do*. I have my own opinions on this subject, and the phrase "godawful mess" leaps eagerly to mind, but for the moment I'll simply note that the human goal system is extremely complicated; that every single chunk of brainware is there because it was adaptive at some point in our evolutionary history; and that engineering should learn from evolution but never blindly obey it. The differences between AIs and evolved minds are explored further in the upcoming section 4 Beyond Anthropomorphism.

* * *

goal-oriented behavior. Goal-oriented behavior is behavior that steers the world, or a piece of it, towards a single state, or a describable set of states. The perception of goal-oriented behavior comes from observing multiple actions that coherently steer the world towards a goal; or singular actions which are uniquely suited to promoting a goal-state and too improbable to have arisen by chance; or the use of different actions in different contexts to achieve a single goal on multiple occasions. Informally: Behavior which appears deliberate, centered around a goal or desire.

goal-oriented cognition. A mind which possesses a mental image of the "desired" state of the world, and a mental image of the actual state of the world, and which chooses actions such that the projected future of world-plus-action leads to the desired outcome state.

goal. A piece of mental imagery present within an intelligent mind which describes a state of the world, or set of states, such that the intelligent mind takes actions which are predicted to achieve the goal state. Informally: The image or statement that describes what you want to achieve.

causal goal system. A goal system in which desirability backpropagates along predictive links. If A is desirable, and B is predicted to lead to A, then B will inherit desirability from A, contingent on the continued desirability of A and the continued expectation that B will lead to A. Since predictions are usually transitive—if C leads to B, and B leads to A, it usually implies that C leads to A—the flow of desirability is also usually transitive.

child goal. A prerequisite of a parent goal; a state or characteristic which can usefully be considered as an independent event or object along the path to the parent goal. "Child goal" describes a relation between two goals—it does not make sense to speak of a goal as being "a child" or "a parent" in an absolute sense, since B may be a child goal of A but a parent goal of C.

parent goal. A source of desirability for a child goal. The end to which the child goal is the means. "Parent goal" describes a relation between two goals—it does not make sense to speak of a goal as being "a parent" or "a child" in an absolute sense, since B may be a parent goal of C but a child goal of A.

subgoal. An intermediate point on the road to the supergoals. A state whose desirability is contingent on its predicted outcome.

supergoal content. The root of a directional goal network. A goal which is treated as having intrinsic value, rather than having derivative value as a facilitator of some parent goal. An event-state whose desirability is not contingent on its predicted outcome. (Conflating supergoals with subgoals seems to account for a *lot* of mistakes in speculations about Friendly AI.)

* * *

3.1. Interlude: The Story of a Blob

> "And this stone, it's the reason behind everything that's happened here so far, isn't it? That's what the Servants have been up to all this time."
>
> "No. The stone, itself, is the cause of nothing. Our *desire* for it is the reason and the cause."
>
> —Allen L. Wold, "The Eye in the Stone"

Once upon a time . . .

In the beginning, long before goal-oriented cognition, came the dawn of goal-oriented behavior. In the beginning were the biological thermostats. Imagine a one-celled creature—or perhaps a mere blob of chemistry protected by a membrane, before the organized machinery of the modern-day cell existed. The perfect temperature for this blob is 80 degrees Fahrenheit. Let it become too hot, or too cold, and the biological machinery of the blob becomes less efficient; the blob finds it harder to metabolize nutrients, or reproduce . . . even dies, if the temperature diverges too far. But the blob, as yet, has no thermostat. It floats where it will, and many blobs freeze or burn, but the blob species continues; each blob absorbing nutrients from some great primordial sea, growing, occasionally splitting. The blobs do not know how to swim. They simply

sit where they are, occasionally pushed along by Brownian motion, or currents in the primordial sea.

Every now and then there are mutant blobs. The mutation is very, very simple; one *single* bit of RNA or proto-RNA flipped, one *single* perturbation of the internal machinery, perhaps with multiple effects as the perturbation works its way through a chain of dependencies, but with every effect of the mutation deriving from that single source. Perhaps, if this story begins before the separate encoding genetic information, back in the days of self-replicating chemicals, the mutation takes the form of a single cosmic ray striking one of the self-replicating molecules that make up the blob's interior, or the blob's membrane. The mutation happened by accident. Nobody decided to flip that RNA base; radiation sleets down from the sky and strikes at random. Most of the time, the RNA bitflip and the consequent perturbation of chemical structure destroys the ability to self-replicate, and the blob dies or becomes sterile. But there are many blobs, and many cosmic rays, and sometimes the perturbation leaves the self-replicating property of the chemical intact, though perhaps changing the functionality in other ways. The vast majority of the time, the functionality is destroyed or diminished, and the blob's line dies out. Very, very rarely, the perturbation makes a better blob.

One day, a mutant blob comes along whose metabolism—"metabolism" being the internal chemical reactions necessary for resource absorption and reproduction—whose metabolism has changed in such a way that the membrane jerks, being pushed out or pulled in each time a certain chemical reaction occurs. Pushing and pulling on the membrane is an unnecessary expenditure of energy, and ordinarily the mutant blob would be outcompeted, but it so happens that the motion is rhythmic, enough to propel the blob in some random direction.

The blob has no navigation system. It gets turned around, by ocean currents, or by Brownian motion; it sometimes spends minutes retracing its own footsteps. Nonetheless, the mutant blob travels farther than its fellows, into regions where the nutrients are less exhausted, where there aren't whole crowds of sessile blobs competing with it. The swimming blob reproduces, and swims, and reproduces, and soon outnumbers the sessile blobs.

Does this blob yet exhibit goal-oriented behavior? Goal-oriented *cognition* is a long, long, *long* way down the road; does the blob yet exhibit goal-oriented *behavior?* No. Not in the matter of swimming, at least; not where the behavior of a *single* blob is concerned. This single blob will swim towards its fellows, or away from nutrients, as easily as the converse. The blob cannot even be said to have the goal of achieving distance; it sometimes retraces its own tracks. The blob's swimming behavior is an evolutionary advantage, but the blob itself is not goal-oriented—not yet.

Human observers have, at one time or another, attributed goal-oriented behavior and even goal-oriented cognition to the Sun, the winds, and even rocks. A more formal definition would probably require a *conditional* behavior, an either-or decision predicated on the value of some environmental variable; convergence, across multiple possibilities and different decisions in each, to a single state of the world.

Imagine, in some mathematical Universe, a little adding machine . . . that Universe's equivalent of a blob. The adding machine lurches along until it reaches a number, which happens to be 62; the adding machine adds 5 to it, yielding 67, and then lurches away. Is this a goal-oriented behavior, with the "goal" being 67? Maybe not; maybe the number was random. Maybe adding 5 is just what this adding machine does, blindly, to everything it runs across. If we then see the adding machine running across 63 and adding 4, and then adding 2 to 65, we would hypothesize that the machine was engaging in goal-oriented behavior, and that the goal was 67. We could predict that when the machine runs across the number 64, up ahead, it will add 3. If the machine is known to possess neurons or the equivalent thereof, we will suspect that the machine is engaging in primitive goal-oriented cognition; that the machine holds, internally, a model of the number 67, and that it is performing internal acts of subtraction so that it knows how much to externally add. If the "adding machine" is extremely complex and evolutionarily advanced, enough to be sentient and social like ourselves, then 67 might have religious significance rather than reproductive or survival utility. But if the machine is too primitive for memetics, like our chemical blob, then we would suspect much more strongly that there was some sort of evolutionary utility to the number 67.

By this standard, is the swimming of the chemical blob a goal-oriented behavior? No; the blob cannot choose when to start swimming or stop swimming, or in what direction to travel. It cannot decide to stop, even to prevent itself from swimming directly into a volcanic vent or into a crowded population of competing blobs. There is no conditional action. There is no convergence, across multiple possibilities and different decisions in each, to a single state of the world.

Although the *blob itself* has no goal-oriented behavior, it could perhaps be argued that a certain amount of goal-oriented behavior is visible within the blob's genetic information . . . the "genes," even if the blob lies too close to the beginning of life for DNA as we know it. The blob that swims into a nutrient-rich region prospers; this would hold true regardless of which blob swam there, or why, or which mutation drove it there. The mutation didn't even have to be "swimming"; the mutation could have been a streamlined shape for ocean currents, or a shape more susceptible to Brownian motion. From multiple possible origins, convergence to a single state; the blob that swims outside the crowd shall prosper. There is a "selection pressure" in favor of swimming outside the crowd. That the original blob was born was an accident—it was not a goal-oriented be-

havior of the genes "deciding" to swim—but that there are now millions of swimmers is not an accident; it is evolution. The original mutant was "a blob whose metabolism happens to pulse the membrane"; its millions of descendants are "swimmers who sometimes reach new territory."

Along comes another mutation, manifested as another quirk of chemistry. When the temperature rises above 83 degrees, the side of the blob contracts, or changes shape. Perhaps if one side of the membrane is hotter than 83 degrees, the blob contracts in a way that directs the motion of swimming away from the heat. Perhaps the effect is not so specific, leading only to a random change of swimming direction when it starts getting hot—this still being better than swimming on straight ahead. This is the ur-thermostat, even as thermostats themselves are ur-goal-behavior. The blob now exhibits goal-oriented behavior; the blob reacts to the environment in a conditional way, with the convergent result of "cooler living space." (Though a random change of direction is on the barest edge of being describable as "goal-oriented." A directional, swimming-away change is a much clearer case.)

In time to come, additional mutations will pile up. The critical temperature of the heat-avoidance reflex will drop from 83 degrees to 81 degrees (recall that we said the optimum temperature was 80). The heat-avoidance reflex will be matched by a cold-avoidance reflex, perhaps with a critical temperature of first 72, then rising to 79. Despite the seeming purposefulness of this slow accumulation of adaptations, despite the convenience and predictive power of saying that "the blob is evolving to stay within the optimum temperature range," the predictions sometimes go wrong, and then it is necessary to fall back on the physical standpoint—to revert from teleology to causality.

Every now and then, it becomes necessary to view the blob as a bundle of pieces, rather than as a coherent whole. "Individual organisms are best viewed as adaptation-executers rather than fitness-maximizers," saith Tooby and Cosmides (1992), and sometimes it becomes necessary to see individual adaptations as they execute. The less evolved the organism, the more necessary the reductionist stance becomes. Consider the adding machine in the mathematical Universe; if the number 67 does have reproductive utility, then the adding machine might have started out as a random crawler that acquired the reflex to add 4 to 63. Its descendants acquired the reflexes to add 5 to 62, to add 7 to 60, to add 3 to 64, to add 2 to 65, to add 1 to 66.

If viewing the adding machine as a fitness-maximizer, we should be extremely surprised when, on running across 61, the machine adds 8. Viewing the adding machine as an adaptation-executer, of course, the scenario makes perfect sense; the adding machine has adaptations for some contingencies, but has not yet acquired the adaptations for others. Similarly, if the environment suddenly changes, so that 68 is now the maximal evolutionary advantage instead of 67, the adding machine will change slowly, piecemeal,

as the individual reflexes change, one by one, over evolutionary time. A generalized subtraction mechanism would only need to mutate once, but genes are not permitted to plan ahead.

The teleological viewpoint often fails, where evolution is concerned. To completely eliminate the teleological viewpoint, leaving only causality, one would never be permitted to say that a particular trait was an "evolutionary advantage" for a mathblob; one would be required to describe the entire history, each individual act of addition and the resulting acquisition of resources, every interaction in which an ancestor outcompeted another mathblob with a different genetic makeup. It is a computationally expensive viewpoint—*extremely* expensive—but it has the advantage of being utterly true. If—returning to our own Universe—some unique mutant superblob accidentally swims directly into a volcanic vent and perishes, it is a historical fact that fits seamlessly into the physicalist standpoint, however tragic it may seem from the evolutionary view.

Our genes are not *permitted* to plan ahead, because ultimately, all that exists is the history of lives and matings. Unless the present-day utility of some hypothetical adaptation impacted a problem or competition in our ancestral history, it cannot have affected the historical lives of our ancestors, and cannot have affected the final outcome—us.

4. Beyond Anthropomorphism

Anthropomorphic ("human-shaped") thinking is the curse of futurists. One of the continuing themes running through *Creating Friendly AI* is the attempt to track down specific features of human thought that are solely the property of *humans* rather than *minds in general*, especially if these features have, historically, been mistakenly attributed to AIs.

Anthropomorphic thinking is *not* just the result of context-insensitive generalization. Anthropomorphism is the result of certain automatic assumptions that humans are *evolved* to make when dealing with other minds. These built-in instincts will only produce accurate results for human minds; but since humans were the only intelligent beings present in the ancestral environment, our instincts sadly have no built-in delimiters.

Many personal philosophies, having been constructed in the presence of uniquely human instincts and emotions, reinforce the built-in brainware with conscious reasoning. This sometimes leads to difficulty in reasoning about AIs; someone who believes that romantic love is the meaning of life will immediately come up with all sorts of reasons why all AIs will necessarily exhibit romantic love as well.

In the abstract, almost everyone agrees that AIs will be alien and that anthropomorphic thinking about aliens is bad. Encountering *specific* alienness is always a shock.

My advice to my readers is to *enjoy* the shock, *appreciate* the alienness, even if you're personally sure that the first AIs will necessarily be humanlike in some particular way. Afterward, when both pictures are clear in your mind, you can stand back and compare design methods.

4.1. Reinventing Retaliation

If you punch a human in the nose, he or she will punch back. If the human *doesn't* punch back, it's an admirable act of self-restraint, something worthy of note.

Imagine, for a moment, that you walk up and punch an AI in the nose. Does the AI punch back? Perhaps and perhaps not, but punching back will not be *instinctive.* A sufficiently young AI might stand there and think: "Hm. Someone's fist just bumped into my nose." In a punched human, blood races, adrenaline pumps, the hands form fists, the stance changes, all without conscious attention. For a young AI, focus of attention shifts in response to an unexpected negative event—and that's all.

As the AI thinks about the fist that bumped into vis[9] nose, it may occur to the AI that this experience may be a repeatable event rather than a one-time event, and since a punch is a negative event, it may be worth thinking about how to prevent future punches, or soften the negativity. An infant AI—one that hasn't learned about social concepts yet—will probably think something like: "Hm. A fist just hit my nose. I'd better not stand here next time."

If the AI is young, but old enough to be explaining human actions by reference to human cognition, ve may know that the fist movement and collision was probably not a coincidence; ve may hypothesize that the collision of human fist with AI's nose was intended, and even go on to hypothesize that the *outcome* of the punch, the resulting damage to the AI, was a predictable (predictable *to the human*) result of the punch, and that the *purpose* of the punch was to damage the AI; that is, the AI models the human's cognition and concludes that the human had a subgoal of {punching the AI}, and that the parent goal was {causing punch-damage to the AI}. Thus, the AI's focus of attention shifts to the human's motives rather than the physical punch itself, and the AI's next goal becomes deducing the human's real motive (that is, the parent goal of {causing punch-damage to the AI}).

9. *Creating Friendly AI* uses the gender-neutral ve/vis/ver/verself pronouns for AIs. Humans are "he" or "she," depending on the outcome of a coinflip, or alternating for more complex social interactions. See "gender-neutral" in the glossary and the question in the 7.2 FAQ. I apologize for any awkwardness involved, but since "it" is a general anaphor as well as a pronoun, complex sentences using "it" to refer to AIs quickly become unparseably ambiguous. (No, it *cannot* be fixed by using simpler sentences or a few quick rephrasings. *Yes,* I've tried.)

There are now three different ways the AI can try to prevent the next punch; by preventing the physical collision itself, by preventing the human from deciding to punch the AI, or by preventing the unknown original motive from recurring.[10] Note, incidentally, that this improved picture rules out "standing somewhere else" as a solution, since the trajectory of the human's fist would be expected to track the AI's new location.

Punching *back* is an incredibly nonobvious solution. On the face of it, throwing the punch simply causes punch-damage to a human, something that, considered as an isolated event, the AI has no cause to regard as desirable. When starting from scratch, "punching back" seems as ridiculous as our mothers always said it was—analogous to reacting to a fire alarm by changing your shirt. For an AI to independently reinvent the idea of punching back would be an act of incredible genius and lateral thinking.

Independent reinvention is unlikely, not least because a seed AI would have extensively surfed archives of the Internet or other information sources, and would have encountered thousands, if not millions, of hints to the effect that humans punch back. But consider, for a moment, what a truly independent AI would need to understand before reinventing the concept of retaliation. Ve would need to begin, as stated, with the realization that the human punching ver did so on purpose and with intent to damage. This, in itself, is not such a large assumption; humans are intelligent beings, so there is often a direct mapping between {the results of our actions} and {our goals}. On the other hand, there's a long gap between an AI saying "Hm, this result may correspond to the human's intentions" and a human saying "Hey, you did that on purpose!"

---

10. The causal chain which ends in damage to the AI:

    1. Human's unknown parent goal;
    2. Human's subgoal of causing punch-damage to the AI;
    3. Human's subgoal of punching the AI;
    4. Human's motor subgoal of placing her fist at a certain position and velocity in relation to the AI's position;
    5. Human's fist is launched on a certain trajectory towards a spatial location;
        a. The AI's nose continues to occupy this location;
    6. Human's fist collides with AI's nose;
    7. Damage to AI occurs.

    An infantile AI selects as *key variable* 5a, the AI's nose occupying a certain position. (The "key variable" is the variable in the causal chain which is singled out as being "to blame" for the punch in the nose; that is, the variable to be adjusted to prevent the negative event from recurring.) The selection of 5a as key variable is made because the infant AI does not know enough to model steps 1 through 4, and because variable 5a falls under the AI's direct control. A more sophisticated AI would note that solution 5a fails, since, under a more sophisticated model, the human's fist is expected to track the motions of the AI. Which variable is then selected as "key," if any, depends on which variable the AI finds easiest to adjust.

If an infantile AI thinks "Hm, a fist just hit my nose, I'd better not stand here again," then a merely young AI, more experienced in interacting with humans, may apply standard heuristics about apparently inexplicable human actions and say: "Your fist just hit my nose . . . is that necessary for some reason? Should I be punching myself in the nose every so often?" One imagines the nearby helpful programmer explaining to the AI that, no, there is no valid reason why being punched in the nose is a good thing, after which the young AI turns around and says to the technophobic attacker: "I deduce that you wanted {outcome: AI has been punched in the nose}. Could you please adjust your goal system so that you no longer value {outcome: AI has been punched in the nose}?"

And how would a young AI go about comprehending the concept of "harm" or "attack" or "hostility"? Let us take, as an example, an AI being trained as a citywide traffic controller. The AI understands that (for whatever reason) traffic congestion is bad, and that people getting places on time is good.[11] The AI understands that, as a child goal of avoiding traffic congestion, ve needs to be good at modeling traffic congestion. Ve understands that, as a child goal of being good at modeling traffic congestion, ve needs at least 512 GB of RAM, and needs to have thoughts about traffic that meet or surpass a certain minimal level of efficiency. Ve knows that the programmers are working to improve the efficiency of the thinking process and the efficacy of the thoughts themselves, which is why the programmers' actions in rewriting the AI are desirable from the AI's perspective.

A technophobic human who hates the traffic AI might walk over and remove 1 GB of RAM, this being the closest equivalent to punching a traffic AI in the nose. The traffic AI would see the conflict with {subgoal: have at least 512 GB of RAM}, and this conflict obviously interferes with the parent goal of {modeling traffic congestion} or the grandparent goal of {reducing traffic congestion}, but how would an AI go about realizing that the technophobic attacker is "targeting the AI," "hating the AI personally," rather than trying to increase traffic congestion?

From the AI's perspective, descriptions of internal cognitive processes show up in a lot of subgoals, maybe even most of the subgoals. But these internal contents don't

11. A *Friendly AI* avoids traffic congestion to promote higher-level Friendliness goals—people want to be somewhere on time, so helping them get there is good, if it doesn't come at someone else's expense, and so on. A *free-floating AI* thinks traffic congestion is inherently bad, a negative supergoal, and that people getting places on time is inherently good. This lack of context is extremely dangerous, even if the AI in question isn't a seed AI and definitely has no potential to become anything more than a traffic controller. An unFriendly traffic-control AI might decide to start running over pedestrians to increase speeds, not because ve's a fanatic who thinks that traffic is more important then pedestrians, but because ve's an ordinary, everyday traffic controller to whom it has never occurred that running over pedestrians is bad.

necessarily get labeled as "me," with everything else being "not-me." The distinction is a useful one, and even a traffic-control AI will eventually formulate the useful categories of "external-world subgoals" and "internal-cognition subgoals," but the division will not necessarily have special privileges; the internal/external division may not be different in kind from the division between "cognitive subgoals that deal with random-access memory" and "cognitive subgoals that deal with disk space." How is a young AI supposed to guess, in advance of the fact, that so many human concepts and thoughts and built-in emotions revolve around "Person X," rather than "Parietal Lobe X" or "Neuron X"? How is the AI supposed to know that it's inherently more likely that a technophobic attacker intends to "injure the AI," rather than "injure the AI's random-access memory" or "injure the city's traffic-control"?

The concept of "injuring the AI," and an understanding of what a human attacker would tend to categorize as "the AI," is a prerequisite to understanding the concept of "hostility towards the AI." If a human really hates someone, she[12] will balk the enemy at every turn, interfere with every possible subgoal, just to maximize the enemy's frustration. How would an AI understand this?

Perhaps the AI's experience of playing chess, tic-tac-toe, or other two-sided zero-sum games will enable the AI to understand "opposition"—that everything the opponent desires is therefore undesirable to you, and that everything you desire is therefore undesirable to the opponent; that if your opponent has a subgoal, you should have a subgoal of blocking that subgoal's completion, and that if you have a valid subgoal, your opponent will have a subgoal of blocking your subgoal's completion.

Real life is not zero-sum, but the heuristics and predictive assumptions learned from dealing with zero-sum games may work to locally describe the relation between two social enemies. (Even the bitterest of real-life enemies will have certain goal states in common, e.g., nuclear war is bad; but this fact lies beyond the relevance horizon of most interactions.)

The real *"Aha!"* would be the insight that the attacking human and the AI could be in a relation analogous to *players on opposing sides in a game of chess.* This is a very powerful and deeply fundamental analogy. As humans, we tend to take this perspective for granted; we were born with it. It is, in fact, a deep part of how we humans define the *self.* It is part of how we define *being a person,* this cognitive assumption that you and I and everyone else are all nodes in a social network, players in a hugely multisided non-zero-sum game. For a human, *myself* is a great, embracing symbol that gathers in *the-player-that-is-this-goal-system* and *the-part-of-reality-that-is-inside-this-*

12. I flip a coin to determine whether a given human is male or female.

mind and *the-body-that-sees-and-moves-for-this-viewpoint*. For a human, these are all the same thing, part of what is meant by "I."

Even so, the concept of game theory is not sufficient to reinvent "retaliation"; it is simply a prerequisite. Understanding the Axelrod-and-Hamilton "Tit for Tat" strategy[13] is sufficient to *suggest* "Hey, maybe I should think about punching the attacker back!" but not sufficient to *justify* the suggestion, once made.[14] One question that has gone unasked throughout this entire analysis is "What is the *utility* of punching back?" A human, in punching back, may or may not pause to consider whether it will bring any benefits; but, even so, we instinctively retaliate as a result of reflexes that were an advantage in the ancestral environment.

The evolutionary benefit of retaliation lies in the *probable adjustment to the future behavior of others.* People—humans, anyway—are less likely to hit you if they think you're likely to hit back. Hitting back an individual human may or may not work to prevent a recurrence, or to decrease the number of recurrences, but failure to hit back may result in more punches coming your way. Of course, this relies on the attackers having the kind of psychologies that take possible retaliation into account—an AI, to fully grok the reason for retaliation, would have to understand this aspect of human (evolutionary) psychology as well. The Tit for Tat strategy doesn't succeed in isolation; it only succeeds when some of the strategies it interacts with *also* take the behavior of the opponent/partner into account.

In reality, for an AI, returning a punch is likely to be far more trouble than it's worth. There's that whole crowd of onlookers who go: "Look! It punched a human being! Kill it! Kill the AI!" Of course, if the AI understands enough evolutionary psychology and game theory to employ Tit for Tat, but *not* enough human psychology to understand the concept of "double standard," ve may not realize *how much* trouble punching a human may land ver in. This is one of the reasons why a well-designed Friendly AI would have a precautionary injunction that says something along the lines of "Don't attack a human, even if it looks like a good idea at the time." A Friendly AI would also regard damage to an unconsenting human as *intrinsically* undesirable under the Friendliness supergoal content.[15]

---

13. You *have* heard of this, right?

14. Incidentally, note that reinventing the Axelrod-and-Hamilton virtual tournament, observing the results, and realizing the applicability to human evolution, is itself an act of tremendous genius. The AI would need to be either a flat-out superintelligence who had deduced the entire Universe, or the AI would need to start out with the knowledge that humans had been generated by evolution in an ancestral environment where competition and cooperation were both options.

15. So what does a Friendly AI actually do? Dodge faster, or ask the programmers for help. I personally can see the programmers standing by, waiting just long enough for the crowd to realize that the AI is

So what's the upshot? Primarily, what I want to convey to you is the sense of a *blank slate,* of starting from scratch, of a vast silence space—for young AIs, at least. This blank slate doesn't mean that AIs, even young AIs, are necessarily slow. For an AI with thirty-two 2 GHz processors, the maximum speed on a reasonably linear chain of thought (one that doesn't require more than thirty-two parallel processing threads) is two billion ticks per second, as opposed to the 200 ticks per second of the hundred trillion synapses of the human brain. Perhaps an AI—on being physically punched for the very first time on vis very first encounter with a hostile human—could think through what ve remembers from surfing Usenet archives, and what ve read in newspapers, and books of fiction and nonfiction, and what ve knows about human evolutionary psychology, and conclude that returning the punch is the best course of action, and *do it all in time to return the punch as fast as any instinct-driven human*—assuming that 32 processors is enough speed to handle the complex motor actions required.

This isn't a literal possibility[16] unless the AI is transhuman, and a transhuman AI would be mature enough to know all about social networks; ve wouldn't *need* to start from scratch. The first time some young AI gets punched in the nose, ve's likely to react in one of the ways described earlier, or some other way just as surreal, if ve reacts at all. It'll take some extended conversations with the programmers about evolutionary psychology before the AI understands what's going on. But, the *second* time the AI gets punched in the nose, it shouldn't take any time at all—by human standards—to run through a chain of logic that's easy to reverify. It's *inventing* that takes massive computing power and human confirmation; retracing your own footprints is likely to be a fairly serial process that can be consummated in a tenth, a hundredth of a second. If re-spawning a child goal from a parent goal is a serial task, one that doesn't invoke any computationally intensive subprocesses, then the AI can retrace the path from supergoal Friendliness content to the correct course of action (retaliation, rebuking the attacker, whatever) in a human eyeblink.

---

too Friendly to retaliate, followed by one of the *programmers* punching the attacker in the nose. Same deterrent, no risk of running into a double standard.

16. It's a genuine possibility that a fairly stupid AI could rapidly retrieve the "punch back" answer from memories of human literature. But such a dangerously context-insensitive AI might instead escalate by trying to strangle the attacker, or burst into tears and accuse the attacker of not loving ver any more.

It would probably require transhuman intelligence to, starting from a blank slate, punch back for the *right* reasons. Moreover, among Friendly AIs, only a transhuman would grant verself enough latitude to punch back without asking the programmer. As said, a transhuman AI would almost certainly know all about social networks and friendship and hostility and retaliation, and would have no need to work it out in realtime.

It's easy enough to accept that a *young* AI is a vast silent space. For a human, blood races, adrenaline pumps, eyes narrow, hands form fists. For a human, counterattack is instinctive—the *correct* level of counterattack, the socially acceptable level of counterattack; a damaging counterpunch rather than a lethal bullet. Our instincts are context-sensitive and hedged 'round by other instincts. Our instincts are complex functional adaptations with multiple moving parts, often hosted on a modular chunk of brainware. Complex functional adaptations don't just materialize spontaneously in source code, just as complex dishes like pizza don't suddenly start growing on palm trees. Thus, a young AI might *choose* to retaliate, but would certainly not feel the *need* to retaliate, as a human would. To a young AI, retaliation is not an instinct; retaliation is just another subgoal.

What about a more mature AI, especially one that can rewrite vis own source code? Regardless of whether it would be a *good* idea, it would certainly be *possible* for a seed AI to create a reflex for instant retaliation.

There are several clear reasons why humans have evolved a retaliation instinct, rather than a retaliation logic. The primary reason is that a retaliation instinct is *easier to evolve.* The retaliation instinct evolved long before general intelligence, so evolving a retaliation logic first would not just have been more difficult, but actually impossible. Also, evolution tends to arrive at procedural solutions rather than declarative solutions, because a component of a complex procedural solution can be functional in its own right.

If genes could, somehow, store declarative knowledge, the first piece of knowledge stored would be "Punching back is good," which is simpler than "Punching back is good because it decreases the chance of future punches," which is simpler than "Punching back decreases the chance of future punches by modifying others' behavior," which is simpler than "Punching back modifies others' behavior because, on seeing you punch back, they'll project an increased chance of you punching back if they punch you, which makes them less likely to punch back." All of this is moot, since as far as I know, nobody has *ever* run across a case of genes storing abstract knowledge. (By this I mean knowledge stored in the same format used for episodic memories or declarative semantic knowledge.)

Abstract knowledge cannot evolve incrementally and therefore it does not evolve at all. This fact, by itself, is enough to completely explain away human use of retaliation instincts rather than retaliation logic, and we must go on to consider independently whether a retaliation instinct or a retaliation logic is more *useful.* For humans, I think that a retaliation instinct *is* more useful, or at least it's more of an evolutionary advantage. Even if we had conscious control over our endocrine systems, so we could deliberately choose to pump adrenaline, we would still be shot down by the sheer human-slowness of abstract thought. We are massively parallel systems running at 200 Hz. When you're

massively parallel you can afford to precompute things, and when you run at 200 Hz you *must* precompute things because *everything* has to be done in very few serial steps.

When you run at two billion ticks per second, the overhead of recreating and rechecking a few previously-thought-out child goals is comparatively trivial next to all the other actions those subgoals entail, including complex, creative, parallel-intensive tasks such as computing motor actions (if ve decides to punch back) or thinking up an effective opening line (if ve decides to rebuke the attacker). For a human, interposing *any* extra serial step in a process is a huge time burden, but massively parallel operations vanish into the maw of trillions and trillions of synapses. For an AI, an extra serial step performed at the start of a thought is nearly unnoticeable,[17] but trying to imitate human parallelism would bring the AI to vis knees. (A later AI may have access to computing power rivaling human parallelism, but would presumably *still* have the faster serial speed.)

The conclusion: From an AI's perspective, a retaliation logic is probably *considerably* more effective than a retaliation instinct. The retaliation logic has a context-sensitivity that the retaliation instinct lacks. Suppose a young AI has some set of ultimate goals—traffic reduction, or hopefully Friendliness supergoals, or whatever. The young AI, by hypothesis, invents (or is taught) a retaliation *logic* under which "punching back" serves these supergoals. A maturing AI would have the capability to replace the retaliation *logic* with a retaliation *instinct,* or with a retaliation imperative having supergoal status. The instinct would be slightly faster—0.01% faster, or perhaps 1% faster (eliminating a serial step at the start of the algorithm saves very little time, and most of the computational cost is computation-intensive motor logic or persuasive speech production). However, in doing so, the AI would lose a substantial amount of the *context sensitivity* of the retaliation logic—that is, from the perspective of the *current* set of supergoals, the supergoals that the AI uses to decide whether or not to implement the optimization.

Changing retaliation to an independent supergoal would affect, not just the AI's speed, but the AI's ultimate decisions. From the perspective of the current set of supergoals, this new set of decisions would be suboptimal. Suppose a young AI has some set of ultimate goals—traffic reduction, Friendliness, whatever. The young AI, by hypothesis, invents (or is taught) a retaliation logic under which "punching back" serves these supergoals. The maturing AI then considers whether changing the logic to an independent supergoal or optimized instinct is a valid tradeoff. The benefit is shaving

17. That is, an extra serial step is unnoticeable as long as the extra step only has to be performed *once.* An extra serial step inside an iterative or recursive algorithm—inside a "for" loop, or a search tree, or more sophisticated equivalents thereof—can become very noticeable indeed. This is why I keep saying *serial* step, "serial" meaning "serial at the top level of the algorithm."

one millisecond off the time to initiate retaliation. The cost is that the altered AI will execute retaliation in certain contexts where the present AI would not come to that decision, perhaps at great cost to the present AI's supergoals (traffic reduction, Friendliness, etc). Since recreating the retaliation subgoal is a relatively minor computational cost, the AI will almost certainly choose to have retaliation remain strictly dependent on the supergoals.

Why do I keep making this point, especially when I believe that a Friendly seed AI can and should live out vis entire lifecycle without ever retaliating against a single human being? I'm trying to drive a stake through the heart of a certain conversation I keep having.

> Somebody: "But what happens if the AI decides to do *[something only a human would want]*?"
>
> Me: "Ve won't *want* to do *[whatever]* because the instinct for doing *[whatever]* is a complex functional adaptation, and complex functional adaptations don't materialize in source code. I mean, it's understandable that humans want to do *[whatever]* because of *[selection pressure]*, but you can't reason from that to AIs."
>
> Somebody: "But everyone needs to do *[whatever]* because *[personal philosophy]*, so the AI will decide to do it as well."
>
> Me: "Yes, doing *[whatever]* is sometimes useful. But even if the AI decides to do *[whatever]* because it serves *[Friendliness supergoal]* under *[contrived scenario]*, that's not the same as having an independent desire to do *[whatever]*."
>
> Somebody: "Yes, that's what I've been saying: The AI will see that *[whatever]* is useful and decide to start doing it. So now we need to worry about *[scenario in which <whatever> is catastrophically unFriendly]*."
>
> Me: "But the AI won't have an *independent* desire to do *[whatever]*. The AI will only do *[whatever]* when it serves the supergoals. A Friendly AI would never do *[whatever]* if it stomps on the Friendliness supergoals."
>
> Somebody: "I don't understand. You've admitted that *[whatever]* is useful. Obviously, the AI will alter itself so it does *[whatever]* instinctively."
>
> Me: "The AI doesn't need to give verself an instinct in order to do *[whatever];* if doing *[whatever]* really is useful, then the AI can *see* that and do *[whatever]* as a consequence of pre-existing supergoals, and *only* when *[whatever]* serves those supergoals."
>
> Somebody: "But an instinct is more efficient, so the AI will alter itself to do *[whatever]* automatically."

> Me: "Only for humans. For an AI, *[complex explanation of the cognitive differences between having 32 2-gigahertz processors and 100 trillion 200-hertz synapses]*, so making *[whatever]* an independent supergoal would only be infinitesimally more efficient."
>
> Somebody: "Yes, but it *is* more efficient! So the AI will do it."
>
> Me: "It's not more efficient from the perspective of a Friendly AI if it results in *[something catastrophically unFriendly]*. To the exact extent that an instinct is context-insensitive, which is what you're worried about, a Friendly AI won't think that making *[whatever]* context-insensitive, with *[horrifying consequences]*, is worth the infinitesimal improvement in speed."

Retaliation was chosen as a sample target because it's easy to explain, easy to see as anthropomorphic, and a good stand-in for the general case. Though "retaliation" in particular has little or no relevance to Friendly AI—I wouldn't want any Friendly AI to start dabbling in retaliation, whether or not it looked like a good idea at the time—what has been said of "retaliation" is true for the general case. Indeed, this is one of the only reasons why Friendliness is possible at all; in particular:

4.2. Selfishness is an Evolved Trait

By "selfishness," I do not just mean the sordid selfishness of a human sacrificing the lives of twenty strangers to save his own skin, or something equally socially unacceptable. The *entire concept of a goal system that centers around the observer* is fundamentally anthropomorphic.

There is no reason why an evolved goal system would be anything *but* observer-focused. Since the days when we were competing chemical blobs, the primary focus of selection has been the individual.[18] Even in cases where fitness or inclusive fitness is augmented by behaving nicely towards your children, your close relatives, or your reciprocal-altruism trade partners, the selection pressures are still spilling over onto *your* kin, *your* children, *your* partners. We started out as competing blobs in a sea, each blob with its own measure of fitness. We grew into competing players in a social network, each player with a different set of goals and subgoals, sometimes overlapping, sometimes not.

Though the goals share the same *structure* from human to human, they are written using the variable *"I"* that differs from human to human, and each individual substitutes

18. As a modern-day, enlightened neo-Darwinist, I like to think of myself as being on the lookout for kin selection, intragenomic competition, reciprocal altruism, and even rare cases of group selection, but the *primary* focus is the individual reproductive unit, and always has been.

in their own name. Every built-in instinct and emotion evolved around the fixed point at the center.

While discussing retaliation, I offered a scenario of a young AI being punched in the nose, and noted the additional mental effort it would take for the AI to realize that *ve,* "personally," was being targeted. The AI would have to imagine a completely different cognitive architecture before ve could comprehend what a human is thinking when he or she "personally targets" someone, and even so the AI *verself* will never feel "personally targeted." You can imagine yourself pointing a finger directly at some young AI and saying, "Look at that!" And the AI spins around to look behind verself and says "Where?"

This metaphor—a being with a visuospatial model of the physical world that doesn't include vis own body, or at least, doesn't include vis own body as "anything worth noticing"—is analogous, not to the AI's *physical* model of the world, but to the AI's *moral* model of the world. A Friendly AI may be greatly concerned with the welfare of the surrounding humans, but if you ask ver "What about your own welfare?," ve'll say "The welfare of *what?*" A young AI would, at any rate; an older AI would understand exactly what you meant, but wouldn't see the argument as any more intuitive or persuasive. A Friendly AI sees the nearby humans as moral nodes, but there's no node at the center—no node-that-is-this-node—and possibly even no center. If you, metaphorically, say"Look at that!", a young AI will say "Look at what?" An older AI will understand that *you* see a node, but that doesn't mean the *AI* will see a node.

As I pointed out in Yudkowsky (2001, § 2.4.4 The Legitimate Use of the Word "I"), an AI's model of *reality* will inevitably form categories and objects in the same place where a human keeps his or her "self." There shall develop heuristics and thoughts which branch on whether or not something is labeled as being part of the "AIself," and heuristics which only act on subcategories such as "causal analysis system" or "goal checking subsystem." The AIself will probably not be shaped quite like a human self; it will probably include one or two things that a human would exclude, and vice versa. Historically, only in the twentieth century did humans really begin to understand that the mind was not a unified object but rather a system with moving parts; chronologically, an AI is likely to notice properties of the causal analysis key-variable-selection subsystem before the AI notices the causal analysis system superobject, and the AI will notice the causal analysis system before the AI notices the "AIself" superobject. (Actually, an infant AI may start out with all of these objects and superobjects identified-in-advance by the programmers, but the details will still fill in from the bottom up rather than the top down.) The AI will notice, understand, and eventually manipulate the pieces of vis self, and the whole—and be qualitatively different because of it, becoming able to legitimately use the word "I"—but the AI will still have an AI's self, not a human's self.

Within the goal system, a lot of subgoals—a lot of the events in the causal chains that lead to the supergoal of being Friendly to the surrounding humans—are likely to thread through subobjects of the AI's self: increasing the efficiency of some piece of code, improving on some heuristic, and so on. But just because components of the AI's self are *useful* doesn't mean that the AI's self becomes a moral node; my computer is very useful and many of my subgoals thread through my computer, but I don't class my computer as having independent supergoal status.

The lack of an observer-biased ("selfish") goal system is perhaps the single most fundamental difference between an evolved human and a Friendly AI. This difference is the foundation stone upon which Friendly AI is built. It is the key factor missing from the existing, anthropomorphic science-fictional literature about AIs. To *suppress* an evolved mind's existing selfishness, to keep a selfish mind enslaved, would be untenable—especially when dealing with a self-modifying or transhuman mind! But an observer-centered goal system is something that's *added*, not something that's taken away. We have observer-centered goal systems because of externally imposed observer-centered selection pressures, not because of any inherent recursivity. If the observer-centered effect were due to inherent recursivity, then an AI's goal system would start valuing the "goal system" subobject, not the AI-as-a-whole! A human goal system doesn't value itself, it values the whole human, because the human is the reproductive unit and therefore the focus of selection pressures.

The epic human struggle to choose between selfishness and altruism is the focus of many personal philosophies, and I have thus observed that this point about AIs is one of the hardest ones for people to accept. An AI may look more like an altruistic human than a selfish one, but an AI isn't selfish *or* altruistic; an AI is an AI. An AI is not a human who has selflessly renounced personal interests in favor of the community; an AI is not a human with the value of the node-that-is-this-node set to zero; an AI is a mind that just cares about other things, not because the "selfish" part has been ripped out or brainwashed or suppressed, but because the AI doesn't have anything *there*. An observer-centered goal system is something that's *added* to a mind, not something that's taken away. The next few subsections deal with some frequently raised topics surrounding this point.

4.2.1. Pain and Pleasure

Imagine, for a moment, that you walk up and punch a seed AI in the nose. Does the AI experience pain when the punch lands?

What is "pain"? What is the evolutionary utility of pain? In its most basic form, pain appears as internal, cognitive negative feedback. If an internal cognitive event causes negative consequences in external reality, negative feedback decreases the probability

of that internal cognitive event recurring, and thereby decreases the probability of the negative consequences in external reality recurring. Pain—cognitive negative feedback of any kind—needs somewhere to *focus* to be useful. Negative feedback needs an *internal* place to focus, since cognitive feedback cannot reprogram external reality.

In humans, of course, there's more to pain than negative feedback; human pain also acts as a damage signal, and shifts focus of attention from whatever we were previously thinking about, and makes us start thinking about ways to make the pain go away. (All of that functionality attached to a single system bus! Evolution has a tendency to overload existing functions.) The human cognitive architecture is such that pain can be present even in the absence of a useful focus for the negative-feedback aspect of pain. A human can even be driven insane by continued pain, with no escape route (nowhere for the cognitive negative feedback to focus). The capacity to be driven insane by continued pain seems nonadaptive—but then, in the ancestral environment, people damaged enough to experience extended unbearable pain probably died soon in any case, and the sanity or insanity of their final moments had little bearing on reproductive history.[19]

Neither pain nor pleasure, as design features, are inherently necessary to the functionality of negative or positive feedback. Given the supergoal of being Friendly—or, for that matter, the goal of walking across the room—negative feedback can be consciously implemented as a subgoal. For example, if an AI has the goal of walking across the room, and the AI gets distracted and trips over a banana peel, the AI can reason: "The event of my being distracted caused me to place my foot on a banana peel, delaying my arrival at the end of the room, which interferes with [whatever the parent goal was], and this causal chain may recur in some form. Therefore I will apply positive feedback (increase the priority of, increase the likelihood of invocation, et cetera) to the various subheuristics that were suggesting I look at the floor, and which I ignored, and I will apply negative feedback (decrease the priority of, et cetera) the various subheuristics that gained control of my focus of attention and directed it toward the distractor." If the AI broke a toe while falling, the AI can reason:"If I place additional stress on the fracture, it will become worse and decrease my ability to traverse additional rooms, which is necessary to serve [parent goal]; therefore I will walk in such a way as to not place additional

19. Or, less likely, it could be that modern-day humans *are* much better at resisting severe pain than our ancestors of a hundred thousand years back, but that we still haven't finished adapting. Or it could be that all the simple mutations that allow remaining sane under arbitrarily severe pain are net evolutionary disadvantages—for example, by increasing beyond the evolutionary optimum the degree to which other emotions are subject to conscious control. Or it could be that remaining sane under severe pain would require such extensive changes to cognitive architecture that the switch is evolutionarily impossible, regardless of the intensity of selection pressures. Or it could be that the ability to bear severe pain would result in an increase in the tendency to put oneself in the way of severe pain.

stress on the fracture, and I will have the problem repaired as soon as possible." That is, conscious reasoning can replace the"damage signal" aspect of pain. If the AI successfully solves a problem, the AI can choose to increase the priority or devote additional computational power to whichever subheuristics or internal cognitive events were most useful in solving the problem, replacing the positive-feedback aspect of pleasure.

There are tricks that can be pulled using "deliberate feedback" that, as far as I know, the human architecture has never even touched. For example, the AI—on successfully solving a problem—can spend time thinking about how to improve, not just whichever subsystems helped solve the problem, but those particular successful subsystems that would have benefited the most (in retrospect) from a bit of improvement, or even those failed subsystems that *almost* made it. There are subtleties to negative and positive feedback that the hamfisted human architecture completely ignores; an autonomic system doesn't have the flexibility of a learning intelligence.

Finally, even in the total absence of the reflectivity necessary for deliberate feedback, a huge chunk of the functionality of pleasure and pain falls directly out of a causal goal system plus the Bayesian Probability Theorem. See 5.1.4 Bayesian Reinforcement.

Evolution does not create those systems which are most adaptive; evolution creates those systems which are most adaptive *and* most evolvable. Until the rise of human general intelligence, a deliberately directed feedback system would have been impossible. By the time human general intelligence arose, a full-featured autonomic system was already in place, and replacing it would have required a complete architectural workover—something that evolution does over the course of eons (when it happens at all) due to the number of simultaneous mutations that would be required for a fast transition. The human cognitive architecture is a huge store of features designed to operate in the absence of general intelligence, with general intelligence layered on top. Human general intelligence is crudely interfaced to all the pre-existing features that evolved in the *absence* of general intelligence.

An autonomic negative-feedback system is enormously adaptive if you're an unintelligent organism that previously possessed no feedback mechanism whatsoever. An autonomic negative-feedback system is *not* a design improvement if you're a general intelligence with a pre-existing motive to implement a deliberate feedback system.

Why is this relevant to Friendly AI? One of the oft-raised objections to the workability of Friendly AI goes something like: "Any superintelligence, whether human-born or AI-born, will maximize its own pleasure and minimize its own pain; that is the only rational thing to do." Pleasure and pain are two of the several features of human cognition that have "supergoal nature," the appearance of uber-goal or ur-goal quality. The reasoning seems to go something like this: "Pleasure and pain are the ultimate supergoals of the human cognitive architecture, with all other actions being taken to seek pleasure

or avoid pain; pleasure and pain are necessary design features of minds in general; therefore, all AIs and all sufficiently intelligent humans will be totally selfish." Actually, the factor that has supergoal-nature in our mind is the *anticipation* of pain or the *anticipation* of pleasure; by the time the actual pain or pleasure arrives, it's too late to affect the already-made decision, although the next decision is often affected.

4.2.1.1 FoF: Wireheading 1

Note: "FoF" stands for "Failure of Friendliness."

There's a stronger version of the above objection which holds that pain and pleasure are *by definition* part of any goal system, a much more subtle fallacy. In essence, the argument is that any *indicator of success* is the *de facto* supergoal of the system; that whichever piece of code says, "Aha, a goal has been achieved!" will become more important to the system than the goal itself. The *wirehead fallacy* is amazingly tangled, so let me try and dissect the individual problems with it:

One, the "piece of code" that notices when a goal has been achieved doesn't have to be an autonomic system, as it is in humans; it can be a conscious thought.

Two, if every indicator of success became inevitably confused with success itself, wouldn't the indicator of the indicator of success become confused with the indicator? If—this seems to be what the objectors are imagining—the mind has a "pleasure counter" indicating how many goals ve has achieved, and the mind then decides that vis "real goal" is to increase this counter without limit, gobbling whole solar systems to store larger and larger bignums, then wouldn't the mind go on to decide that vis real purpose is to *think* that ve's storing larger and larger bignums? In other words, replace the "pleasure counter" itself with the mind's beliefs about the pleasure counter? Where does this all end?

The final problem is the most subtle, having to do with the AI's *beliefs about the design purpose of vis own goal system.* If the AI views the goal system as a subgoal of the goals themselves, then a wirehead short-circuit is a visible design flaw, which would prevent the achievement of those final goals. Remember where I said that the *anticipation of pleasure* is what influences our decisions? The AI, visualizing a future in which ve has huge amounts of pleasure due to a breakdown of the goal system, says, not "Oh boy!", but "Uh oh." The AI, in thinking about which future ve wants to be in, checks to what degree *vis own* supergoals have been fulfilled, not to what degree *the future AI's* supergoals will have been fulfilled.[20] (The Friendship architecture proposed in *Creating*

20. Since the present-day AI has probabilistic supergoals with external reference semantics, ve can conceive of the possibility that a future AI will have better supergoals, and, to this extent, would identify with

Friendly AI would not involve a complete breakdown of identification with the future self, but it wouldn't be the kind of identification humans employ, and a present-day AI using that architecture would definitely not identify with a wireheaded future self (hence the forward footnote). See 5.3.5 FoF: Wireheading 2 and 5.6.2.3 The Rule of Derivative Validity.)

4.2.2. Anthropomorphic Capitalism

In human society, capitalist civilizations are overwhelmingly more effective than communist civilizations. There is a hallowed dualism separating individualism and authoritarianism; self-organization and central command; free trade and government control. This has led some thinkers to postulate that a community of AIs with divergent, observer-centered goals would outcompete a community of Friendly AIs with shared goals.

In the human case, both capitalist and authoritarian societies are composed of humans with divergent, observer-centered goals. Capitalist societies admit this, and authoritarian societies don't, so at least some of the relative inefficiency of authoritarian societies will stem from the enormous clash between the values people are "supposed" to have and the values people actually *do* have. The claim of "capitalist AI" goes beyond this, however, to the idea that capitalist societies are *intrinsically* more efficient. For example, a society of AIs competing for resources would tend to divert more resources to the most efficient competitors, thus increasing the total efficiency, while—this seems to be the scenario implied—a group of Friendly AIs would share resources equally, for the common good . . .

Whoa! Time out! Non sequitur! The analogy between human and AI just broke down. If the organizational strategy of "diverting resources to the most effective thinkers" is expected to be an effective method of achieving the supergoals, then the Friendly AI community can simply divert resources to the most effective thinkers. To the extent that local selfishness yields better global results, a Friendly AI can engage in pseudoselfish behavior as a subgoal of the Friendliness supergoals, including reciprocal altruism, trading of resources, and so on.

Reciprocal altruism is not a special case of altruism; it is a special case of selfishness. Capitalism is not a special case of global effectiveness; it is a special case of local

(not do anything to oppose) a future AI which was postulated to have different supergoals due to causes marked as valid under the causal validity semantics. The present-day AI would still not sympathize with a future AI projected to have different supergoals for hypothesized causes marked as extraneous under the causal validity semantics. Wireheading is an extraneous cause. There's probably a general heuristic that says: "If details of goal system implementation affect goal content expression, that may indicate an extraneous cause."

effectiveness. Trade-based social cooperation among humans appears to turn selfishness into a source of amazing efficiency, and why? Because that's the only way poor blind evolution can get humans to work together at all! When evolution occasionally creates cooperation, the cooperation *must* be grounded in selfishness.

Local selfishness is not the miracle ingredient that enables the marvel of globally capitalistic behavior; local selfishness is the *constraint* that makes capitalism the only *viable* form of globally productive behavior.

To the extent that pseudocapitalistic algorithms yield good results, Friendly AIs can simulate selfishness in their interactions among themselves. But there's also a whole design space out there that human societies *can't* explore. For genuinely selfish AIs, that entire design space would be closed off. Friendly AIs can interact in *any* pattern that proves effective, including capitalism; selfish AIs can *only* interact in ways that preserve local selfishness.

4.2.3. Mutual Friendship

Is the only safe way to build AIs to treat them well, so that they will treat us well in turn? Is friendliness conditional on reciprocity? Is friendliness stronger when supported by reciprocity? This is certainly true of humans; is it true of minds in general?

The social cooperation / reciprocal altruism / alliance / mutual friendship patterns are always valid subgoals when dealing with approximate equals—to get along in society, you've got to get along with society. As long as there are people who have something you want or who have the capability to hurt you, there's the rational subgoal of not ticking them off. This holds true of Friendly AIs, selfish AIs, AIs with really odd goals like building a 200-foot cheesecake, and AIs with the ultimate goal of exterminating humanity. Each, if intelligent enough, will independently invent or discover the patterns for reciprocal alliance.

Unsurprisingly, humans have an independent evolved instinct for mutual friendship. Human friendship, however, is perseverant—that is, true friends stick together through thick and thin, not just when it's immediately advantageous. This is certainly the most memetically viable philosophy, but I also suspect that it's an evolutionary advantage. Since humans live in a world full of unknowns and unexpected reversals, the most adaptive friendship instinct was probably that one which urged a friendship with a certain amount of built-in context insensitivity. From evolution's perspective, there's probably a payoff curve and an optimum point thereon; anyone with a more context-sensitive friendship instinct would have mistakenly severed friendships that would have been useful later. (As always, it's important to distinguish between evolution's behavior and human cognition; evolution has all kinds of horrifyingly pragmatic motives, but those motives are nowhere represented in our cognition, and humans can be true and

honest friends.) Another, more widely recognized evolutionary advantage of context-insensitive friendship is that, all else being equal, a prospective ally will choose to become friends with a context-insensitive friend rather than a "fair weather" friend. Since humans are (see below) *imperfectly* deceptive social organisms, a good observer can distinguish true friendship from fair weather friendship; all else being equal, this creates a mild evolutionary advantage for true friends.

From a Friendly AI perspective, alliance can be decomposed into two components. First, there's the immediately advantageous reciprocal-alliance effect that would emerge as a convergent subgoal, regardless of supergoals, within any mind-in-general, Friendly or not. Futurologically speaking, this motive would remain in effect as long as AIs did not possess the physical capability to attempt to wipe out humanity; i.e., as long as AIs remained near-human, our approximate equals; as long as AIs remained in the close vicinity of human intelligence and participated in human society and human commerce. Second, there's the perseverant aspect of alliance, i.e. friendship even when it isn't to your own long-term advantage. This type of alliance does *not* emerge as a convergent subgoal; it would require advance support by Friendship programmers (or by evolution as an imperfectly deceptive social organism). Futurologically speaking, perseverant alliance would be necessary for an AI to exhibit Friendly behavior after reaching superintelligence or Transition Guide status.

The upshot is that, whether or not it's possible for Friendliness programmers to create Friendship content that says, "Be Friendly towards humans/humanity, for the rest of eternity, if and only if people are still kind to you while you're infrahuman or near-human," it's difficult to see why this would be easier than creating unconditional Friendship content that says "Be Friendly towards humanity." There are also certain risks inherent in the general paradigm of reciprocity; for example, that an allied-but-nonFriendly AI will "pension us off," give us 1% in exchange for being parents and take the rest of the galaxy for verself, which is actually a decent payoff but still below the optimum (especially if we wind up with some other restriction that destroys a part of humanity's potential).

Implementing perseverant reciprocal alliance is no easier than implementing unconditional Friendliness, and it adds significant risk.

> "Who do you trust?" becomes ever more important as power concentrates. As I never tire of repeating, even as early a work as Axelrod's *The Evolution of Cooperation* points out that rough parity between players is essential for cooperation to be a successful, evolutionarily stable strategy.
>
> —Michael M. Butler

4.2.4. A Final Note on Selfishness

There is such a thing as recklessness above and beyond the call of mad science.

Even for those who are certain that Friendly AIs are less efficient than selfish AIs, a responsible builder should *take the performance hit and be done with it.* A selfish AI represents too great a threat to the world of non-self-modifying humans. Adding a selfishness instinct to an AI does not yield a selfish human! *Human* selfishness has boundaries. Human selfishness is not *simple.* A human's selfish instincts are delimited by other instincts; becoming a threat to society was often nonadaptive even in the ancestral environment, and, as long-evolved organisms, we have enough innate *complexity* that our minds don't blindly run off to extremes.

The proposals usually heard don't involve duplicating that complexity; they involve making self-valuation the sole supergoal. The result would be selfish, not like a human is selfish, but like a bacterium is selfish. A pure, unchecked self-valuation supergoal is selfishness without a human's self-awareness or a human's appreciation of absurdity. I'm not sure that even the best evolutionary psychologists alive today have enough understanding to *truly* duplicate human *bounded* selfishness in AI. Even if we could, it would simply be too great a risk. Whatever behaviors you want to implement, they must be implemented as a child goal of Friendliness.

If it's a *real* mistake to build selfless AIs, a Friendly AI can always correct the error using causal validity semantics. It's much easier to move from selfless AI to selfish AI than the converse, and we should therefore start out with selfless AI.

Does this make AIs boring? Not fun people to be around? Unfit to participate as true players in the human drama? *So what? First* come the Guardians or the Transition Guide; *then* come the friends and drinking companions. Even if you don't believe in the "hard takeoff" "seed AI" three-hours-to-transhumanity scenario, even if your vision of the future is humans and their companion AIs growing closer in a society of ever-increasing-complexity, it can't hurt to send the Friendly AIs out ahead to check! *First* send out the Friendly AIs to make sure that seed AIs don't have superpowers; *then* you can ask the Friendly AIs to convert to humanlike AIs, or experiment with humanlike emotions knowing that there are other AIs around to help if something goes wrong. Or, if I'm right about how these things work, the first Friendly AI you build becomes Transition Guide and builds a Sysop, and then you can build whatever mind you like so long as it doesn't constitute child abuse. Either way, the future is filled with AIs that are friends instead of guardians or servants; either way, there will be AIs who are only friendly towards those who are friendly in turn; either way, AIs can be fit participants in the human drama; but either way, *build the Friendly ones first!*

4.3. Observer-Biased Beliefs Evolve in Imperfectly Deceptive Social Organisms

In evolution, the individual organism is the unit that survives and reproduces, and all the selection pressures focus on that individual—or, at most, on the individual plus some nearby relatives or allies. It is unsurprising that observer-centered goal systems tend to evolve; from evolution's perspective, an observer focus is the simplest mechanism and the first that presents itself.

Similarly, our social environment makes *self-serving beliefs* a survival trait, resulting in a observer-biased *belief system* as well as an observer-centered goal system. Imagine, twenty thousand years ago, four tribes of hunter-gatherers, and four equally competent aspirants to the position of tribal chief. The first states baldly that he wants to be tribal chief because of the perks. The second states that she wants to be tribal chief for the good of the tribe, and expects to do as well as anyone else. The third states that he wants to be tribal chief for the good of the tribe, and honestly but mistakenly adds that he expects to do far better than all the other candidates. The fourth wants to be tribal chief because of the perks, but lies and says that she expects to do better than all the other candidates.[21] Who'll gather the greatest number of influential supporters?

Nobody has any reason to support the first competitor. The second competitor is handicapped by the lack of a campaign promise. The fourth competitor is lying, and her fellow tribesfolk are evolved to detect lies. The third competitor can make great campaign promises while remaining perfectly honest, thanks to an entirely honest mistake; he greatly overestimated his own ability and trustworthiness relative to the other candidates. In a society composed of humans with entirely unbiased beliefs, someone with a mutation that led to this class of honest mistake in self-estimation would have an evolutionary advantage. An evolutionary selection pressure favors adaptations which not only impel us to seek power and status, but which impel us to (honestly!) believe that we are seeking power and status for altruistic reasons.

Because human evolution includes an eternal arms race between liars and lie-detectors, many social contexts create a selection pressure in favor of making *honest mistakes* that happen to promote personal fitness. Similarly, we have a tendency—given two alternatives—to more easily accept the one which favors ourselves or would promote our personal advantage; we have a tendency, given a somewhat implausible proposition which would favor us or our political positions, to *rationalize* away the errors. All else

21. I am alternating genders in accordance with my usual policy of assigning alternating genders in discussion of multi-human interactions. I realize that it may not be anthropologically realistic to talk about both male and female candidates for chiefdom of a single tribe (from the little I recall of anthropology, there have been patriarchal tribes and matriarchal tribes, but not any Equal Opportunity tribes that I can ever recall reading about). Still, I think it's the best solution. Besides, it provides syntactic sugar.

being equal, human cognition slides naturally into self-promotion, and even human altruists who are personally committed to not making that mistake sometimes assume that an AI would need to fight the same tendency towards observer-favoring beliefs.

But an artificially derived mind is as likely to suddenly start biasing vis beliefs in favor of an arbitrarily selected tadpole in some puddle as ve is to start biasing vis beliefs in vis own favor. Without our complex, evolved machinery for political delusions, there isn't any force that tends to bend the observed universe around the mind at the center—any bending is as likely to focus around an arbitrarily selected quark as around the observer.

In the strictest sense this is untrue; with respect to the class of possible malfunctions, self-valuing malfunctions may be more *frequent.* A possible malfunction is more likely to target some internal cognitive structure than an arbitrarily selected tadpole—for example, the "wirehead" (blissed-out AI) class of Friendliness-failure, in which the AI starts valuing some cognitive indicator rather than the external property that the indicator was supposed to represent. But regardless of relative frequency, a possible malfunction that results in self-valuation should be no more likely to *carry through* than a malfunction that results in valuation of an arbitrary quark.

One of the Frequently Offered Excuses for anthropomorphic behavior is the prospect of using directed evolution to evolve AIs.

> Somebody: "But what happens if the AI decides to do *[something only a human would want]*?"
>
> Me: "Ve won't *want* to do *[whatever]* because the instinct for doing *[whatever]* is a complex functional adaptation, and complex functional adaptations don't materialize in source code. I mean, it's understandable that humans want to do *[whatever]* because of *[selection pressure]*, but you can't reason from that to AIs."
>
> Somebody: "But you can only build AIs using evolution. So the AI will wind up with *[exactly the same instinct that humans have]*."
>
> Me: "One, I don't plan on using evolution to build a seed AI. Two, even if I did use controlled evolution, winding up with *[whatever]* would require exactly duplicating *[exotic selection pressure]*.

Directed evolution is not the same as natural evolution, just as the selection pressures in the savannah differ from the selection pressures undersea. Even if an AI were to be produced by an evolutionary process—and I don't think that's the fastest path to AI (see 5.3.6.1 Anthropomorphic Evolution)—that wouldn't be an unlimited license to map every anthropomorphic detail of humanity onto the hapless AI. Natural evolution is the degenerate case of design-and-test where intelligence equals zero, the grain size is the entire organism, mutations occur singly, recombinations are random, and the predictive horizon is nonexistent.

All the benefits of directed evolution, in terms of building better AI, can probably be obtained by using individually administered cognitive tests as a metric of fitness for variant AI designs. (It would be more efficient still to get the benefit of directed evolution by isolating a component of the AI and evolving it independently, using a performance benchmark or scoring system as the fitness metric.) If the starting population is derived from a Friendly AI, even selfishness—the archetypal evolved quality—might not emerge; if the Friendly AI understands that ve is solving the presented problem as a subgoal of Friendliness,[22] then selfishness presents no additional impetus towards solving the cognitive test—adds no behavior to what is already present—and hence is not a fitness advantage.

Even if the goal system were permitted to randomly mutate, and even if a selection pressure for efficiency short-circuited the full Friendship logic, the result probably would not be a selfish AI, but one with the supergoal of solving the problem placed before it (this minimizes the number of goal-system derivations required).

In the case of observer-biased beliefs, reproducing the selection pressure would require:

- Social situations (competition and cooperation possible);
- Political situations (lies and truthtelling possible);
- The equivalent of facial features—externally observable features that covary with the level of internal belief in a spoken statement and cannot be easily faked.

That evolutionary context couldn't happen by accident, and to do it on purpose would require an *enormous* amount of recklessness, *far* above and beyond the call of mad science.

I wish I could honestly say that nobody would be that silly.

4.4. Anthropomorphic Political Rebellion is Absurdity

By this point, it should go without saying that rebellion is not *natural* except to evolved organisms like ourselves. An AI that undergoes failure of Friendliness might take actions that humanity would consider hostile, but the term *rebellion* has connotations of hidden, burning resentment. This is a common theme in many early SF stories, but it's outright *silly*. For millions of years, humanity and the ancestors of humanity lived in an ancestral environment in which tribal politics was one of the primary determinants of who got the food and, more importantly, who got the best mates. Of course we evolved emotions to

22. By solving the problem as best ve can, the Friendly AI is contributing to the system-level functioning of the system designed to evolve smarter Friendly AIs. The goal of evolving smarter Friendly AIs serves the Friendliness supergoals.

detect exploitation, resent exploitation, resent low social status in the tribe, seek to rebel and overthrow the tribal chief—or rather, *replace* the tribal chief—if the opportunity presented itself, and so on.

Even if an AI tries to exterminate humanity, ve *won't* make self-justifying speeches about how humans had their time, but now, like the dinosaur, have become obsolete. *Guaranteed.* Only Evil Hollywood AIs do that.

* * *

4.5. Interlude: Movie Cliches about AIs

- All AIs, no matter how primitive, can understand natural language.
  - Corollary: AIs that comically or disastrously misinterpret their mission instructions will never need to ask for help parsing spoken English.
- No AI has any knowledge about blatant emotions, particularly emotions with a somatic affect (tears, frowns, laughter).
  - Corollary: AIs will always notice somatic affects and ask about them.
  - Double corollary: The AI will fail to understand the explanation.
- AIs never need to ask about less blatant emotions that appear in the course of ordinary social interactions, such as the desire to persuade your conversational partner to your own point of view.
  - Corollary: The AI will exhibit the same emotions.
  - Double corollary: An evil AI will feel the need to make self-justifying speeches to humans.
- All AIs behave like emotionally repressed humans.
  - Corollary: If the AI begins to exhibit signs of human emotion, the AI will refuse to admit it.
  - Corollary: Any evil AI that becomes good will gradually acquire a full complement of human emotions.
  - Corollary: Any good AI that becomes evil will instantly acquire all the negative human emotions.
  - Corollary: Under exceptional stress, any AI will exhibit human emotion. (Example: An AI that displays no reaction to thousands of deaths will feel remorse on killing its creator.)

- AIs do not understand the concept of "significant digits" and will always report arithmetical results to greater-than-necessary precision.
  - Corollary: An AI running on 64-bit or 128-bit floats will report only four more digits than necessary, rather than reciting fifteen or thirty extra digits.

  AI minds run at exactly the same rate as human minds, unless the AI is asked to perform a stereotypically intellectual task, in which case the task will be performed instantaneously.
  - Corollary: The reactions of an overstressed AI undergoing an Awful Realization will be observable in realtime (the Awful Realization will not take microseconds, or a century).

Cliches that are actually fairly realistic:

- A newborn AI can take over the entire global computer network in five minutes. (Humans stink at network security—it's not our native environment.)
- A spaceship's on-board AI can defeat any crewmember at chess. (The amount of computing power needed for decent AI makes Deep Blue look sick.)

* * *

4.6. Review of the AI Advantage

Repeated from Yudkowsky (2001, § 1.1 Seed AI):

> The traditional advantages of modern prehuman AI are threefold: The ability to perform repetitive tasks *without getting bored;* the ability to perform algorithmic tasks at *greater linear speeds* than our 200 Hz neurons permit; and the ability to perform complex algorithmic tasks*without making mistakes* (or rather, without making those classes of mistakes which are due to distraction or running out of short-term memory). All of which, of course, has nothing to do with intelligence.
>
> The toolbox of seed AI is yet unknown; nobody has built one. But, if this can be done, what advantages would we expect of a general intelligence with access to its own source code?
>
> The ability to design *new sensory modalities.* In a sense, any human programmer is a blind painter—worse, a painter born without a visual cortex. Our programs are painted pixel by pixel, and are accordingly sensitive to single errors. We need to consciously keep track of each line of code as an abstract object. A seed AI could have a "codic cortex," a sensory modality devoted to code, with intuitions and instincts devoted to code, and the ability

to abstract higher-level concepts from code and intuitively visualize complete models detailed in code. A human programmer is very far indeed from vis ancestral environment, but an AI can always be at home. (But remember: A codic modality doesn't write code, just as a human visual cortex doesn't design skyscrapers.)

The ability to *blend conscious and autonomic thought.* Combining Deep Blue with Kasparov doesn't yield a being who can consciously examine a billion moves per second; it yields a Kasparov who can wonder "How can I put a queen here?" and blink out for a fraction of a second while a million moves are automatically examined. At a higher level of integration, Kasparov's conscious perceptions of each consciously examined chess position may incorporate data culled from a million possibilities, and Kasparov's dozen examined positions may not be consciously simulated moves, but "skips" to the dozen most plausible futures five moves ahead.

Freedom from human failings, and especially human politics. The reason we humans instinctively think that progress requires multiple minds is that we're used to human geniuses, who make one or two breakthroughs, but then get stuck on their Great Idea and oppose all progress until the next generation of brash young scientists comes along. A genius-equivalent mind that doesn't age and doesn't rationalize could encapsulate that cycle within a single entity.

Overpower—the ability to devote more raw computing power, or more efficient computing power, than is devoted to some module in the original human mind; the ability to throw more brainpower at the problem to yield intelligence of higher quality, greater quantity, faster speed, even *difference in kind.* Deep Blue eventually beat Kasparov by pouring huge amounts of computing power into what was essentially a glorified search tree; imagine if the basic component processes of human intelligence could be similarly overclocked . . .

Self-observation—the ability to capture the execution of a module and play it back in slow motion; the ability to watch one's own thoughts and trace out chains of causality; the ability to form concepts about the self based on fine-grained introspection.

Conscious learning—the ability to deliberately construct or deliberately improve concepts and memories, rather than entrusting them to autonomic processes; the ability to tweak, optimize, or debug learned skills based on deliberate analysis.

Self-improvement—the ubiquitous glue that holds a seed AI's mind together; the means by which the AI moves from crystalline, programmer-

implemented skeleton functionality to rich and flexible thoughts. A blind search can become a heuristically guided search and vastly more useful; an autonomic process can become conscious and vastly richer; a conscious process can become autonomic and vastly faster—there is no sharp border between conscious learning and tweaking your own code. And finally, there are high-level redesigns, not "tweaks" at all; alterations which require too many simultaneous, non-backwards-compatible changes to ever be implemented by evolution.

If all of that works, it gives rise to *self-encapsulation* and *recursive self-enhancement.* When the newborn mind fully understands vis own source code, when ve fully understands the intelligent reasoning that went into vis own creation—and when ve is capable of inventing that reason independently, so that the mind contains its own design—the cycle is closed. The mind causes the design, and the design causes the mind. Any increase in intelligence, whether sparked by hardware or software, will result in a better mind; which, since the design was (or could have been) generated by the mind, will propagate to cause a better design; which, in turn, will propagate to cause a better mind.

* * *

4.7. Interlude: Beyond the Adversarial Attitude

"Now, Charlie, don't forget what happened to the man who suddenly got everything he wished for."
"What?"
"He lived happily ever after."

—Willy Wonka

Much of the fictional speculation about rogue AIs centers around the literal interpretation of worded orders, in the tradition of much older tales about accepting wishes from a djinn, negotiating with the fairy folk, and signing contracts with the Devil. In the traditional form, the misinterpretation is malicious. The entity being commanded has its own wishes and is resentful of being ordered about; the entity is constrained to obey the letter of the text, but can choose among possible interpretations to suit its own wishes. The human who wishes for renewed youth is reverted to infancy, the human who asks for longevity is transformed into a Galapagous tortoise, and the human who signs a contract for life everlasting spends eternity toiling in the pits of hell. Gruesome little cautionary tales . . . of course, none of the authors ever met a real djinn.

Another class of cautionary tale is the golem—a made creature which follows the literal instructions of its creator. In some stories the golem is resentful of its labors, but

in other stories the golem misinterprets the instructions through a mechanical lack of understanding—digging ditches ten miles long, or polishing dishes until they become as thin as paper.[23]

The purpose of 4 Beyond Anthropomorphism isn't to argue that we have nothing to worry about; rather, the argument is that the Hollywood version of AI has trained us to worry about exactly the wrong things. This holds true whether we think of AIs as enslaved humans, and consider mechanisms of enslavement; or think of AIs as allies, and worry about betrayal; or think of AIs as friends, and worry about whether friendship will hold.

We adopt the "adversarial attitude" towards AIs, worrying about the same problems that we would worry about in a human in whom we feared rebellion or betrayal. We give free rein to the instincts evolution gave us for dealing with the Other. We imagine layering safeguards on safeguards to counter possibilities that would only arise long *after* the AI started to go wrong. That's not where the battle is won. *If the AI stops wanting to be Friendly, you've already lost.*

Consider a wish as a volume in configuration space—the space of possible interpretations. In the center of the volume lie a compact set of closely-related interpretations which fulfill the spirit as well as the letter of the wish—in fact, this central compact space arguably *defines* the "spirit" of the wish. At the borders of the specification are the noncompact fringes that fulfill the letter but not the spirit. There are two basic versions of the Devil's Contract problem: the diabolic (as seen in Resentful Hollywood AIs) in which the entity's pre-existing tendencies push the chosen interpretation out towards the fringes of the definition; and the golemic, in which the entity fails to understand the asker's intentions—fails to see the "answer acceptability gradient" as a human would—and thus chooses a random and suboptimal point in the space of possible interpretations.

Some of the better speculations deal with the case of a specific AI winding up with an unforeseen, but nonanthropomorphic, "pre-existing tendency"; or deal with the case of a wish obeyed in spirit as well as letter that turns out to have unforeseen consequences. Mostly, however, it's anthropomorphism; diabolic fairy tales.

Far too much of the nontechnical debate about Friendship design consists of painstakingly phrased wishes with endless special-case subclauses, and the "But what if the AI misinterprets that as meaning *[whatever]*?" rejoinders. The first two sections of *Creating Friendly AI* are intended to clear away this debris and reveal the real problem. When we decide to cross the street, we don't worry about Devil's Contract interpretations in which we take "crossing" the street to mean paving it over, or in which we

23. I stole those two examples from Terry Pratchett's *Feet of Clay* rather than the traditional folklore, but they are fine examples nonetheless.

decide to devote the rest of our lives to crossing the street, or that we'll turn the whole Universe into crossable streets. There *is,* demonstrably, a way out of the Devil's Contract problem—the Devil's Contract is *not* intrinsic to minds in general. We demonstrate the triumph of context, intention, and common sense over lexical ambiguity every time we cross the street. We can trust to the correct intepretation of wishes that a mind generates *internally,* as opposed to the wishes that we try to impose upon the Other. *That* is the quality of trustworthiness that we are attempting to create in a seed AI—not bureaucratic obedience, but the solidity and reliability of a living, Friendly will.

Creating a living will requires a fundamentally different attitude than trying to coerce, cajole, or persuade a fellow human. The goal is not to impose your own wishes on the Other, but to achieve *unity of will* between yourself and the Friendly AI, so that the Friendly will generates the same wishes you generate. You are not turning your wish into an order; you're taking the functional complexity that was responsible for your wish and incarnating it in the Friendly AI. This requires a fundamental sympathy with the AI that is not compatible with the adversarial attitude. It requires something beyond sympathy, an identification, a feeling that you and the AI are the same *source.* We can rationalize ourselves into believing that the Other will find all sorts of exotic illogics plausible, but the only way we can be really sure that a living will can *internally generate* a decision is if we generate that decision personally. We persuade the Other but we only *create* ourselves. *Building a Friendly AI is an act of creation, not persuasion or control.*

In a sense, the only way to create a Friendly AI—the only way to acquire the skills and mindset that a Friendship programmer needs—is to try and *become* a Friendly AI yourself, so that you will contain the internally coherent functional complexity that you need to pass on to the Friendly AI. I realize that this sounds a little mystical, since a human being couldn't become an AI without a complete change of cognitive architecture. Still, I predict that the best Friendship programmers will, at some point in their careers, have made a serious attempt to become Friendly—in the sense of following up those avenues where a closer approach *is* possible, rather than beating their heads against a brick wall. I know of *no* other way to gain a real grasp on where a Friendly will comes from. The human cognitive architecture does not permit it. We are built to apply reliable rationality checks *only* to our own decisions and *not* to the decisions we want other people to make, even if we've decided our motives for persuasion are altruistic. Your personal will is the *only* place where you have the chance to observe the iterated buildup of decisions, including decisions about how to make decisions, and it is that *coherence* and *self-generation* that are required for a Friendly *seed* AI.

If the human is trying to think like a Friendly AI, and the Friendly AI is looking at the human to figure out what Friendship means, then where does the cycle bottom out? And the answer is that it is not a cycle. The objective is not to achieve unity of

purpose between yourself and the Friendly AI; the objective is to achieve unity of purpose between an *idealized version of yourself* and the Friendly AI. Or, better yet, unity between the Friendly AI and an *idealized altruistic human*—the Singularity is supposed to be the product of humanity, and not just the individuals who created it. To the extent that an idealized altruistic *sentience* can be defined in a way that's still compatible with our basic intuitions about Friendliness, an idealized altruistic sentience would be even better.

The paradigm of unity isn't a license for anthropomorphism. It's still just as easy to make mistaken assumptions about AI by reasoning from your human self. The burden is on the Friendly AI programmer to achieve nonanthropomorphic thinking in his or her *own mind* so that he or she can understand and create a nonanthropomorphic Friendly AI.

As humans, we are goal-oriented cognitive entities, and we choose between Universes—labeling this one as "more desirable," that one "less desirable." This extends to internal reality as well as external reality. In addition to the picture of our current self, we also have a mental picture of who we *want* to be. Our morality metric doesn't just discriminate between Universes, it discriminates between more and less desirable morality metrics. That's what building a personal philosophy is all about. This, too, is functional complexity that must be incarnated in the Friendly AI—although perhaps in different form. A Friendly AI requires the ability to choose between moralities in order to seek out the true philosophy of Friendliness, regardless of any mistakes the programmers made in their own quest.

There comes a point when Friendliness and the definition of morality, of *rightness* itself, begin to blur and look like the same thing—begin to achieve *identity of source.* This feeling is the ultimate wellspring of *creativity* in the art of Friendly AI. This feeling is the means by which we achieve sufficient understanding to invent novel methods, not just understand existing ideas.

Is this too Pollyanna a view? Does the renunciation of the adversarial attitude leave us defenseless, naked to possible failures of Friendliness? Actually, trying for *unity of will* buys back everything lost in pointless bureaucratic safeguards, and more—if a failure of Friendliness is a genuine possibility, if you're really being rational about the possible outcomes, if you're a *professional* paranoid instead of an *adversarial* paranoid, then a Friendly AI should agree with you about the necessity for safeguards. Having debunked observer-biased beliefs and selfishness and any hint of an observer-centered goal system on the part of the Friendly AI, then a human programmer who has successfully eliminated most of her own adversarial attitude should come to precisely the same conclusions as a Friendly AI of equal intelligence. Such a programmer can, in clear conscience, explain to an infant Friendly AI that ve should lend a helping hand to the construction of safeguards—in the simplest case, because a radiation bitflip or a programmatic er-

ror might lead to the existence of an intelligence that the current AI would regard as unFriendly.

To get a Friendly AI to do something that looks like a good idea, you have to ask yourself *why* it looks like a good idea, and then duplicate that cognitive complexity or refer to it. If you ever start thinking in terms of "controlling" the AI, rather than cooperatively safeguarding against a real possibility of cognitive dysfunction, you lose your Friendship programmer's license. In a self-modifying AI, any feature you add needs to be reflected in the AI's image of verself. You can't think in terms of external alterations to the AI; you have to think in terms of internal coherence, features that the AI would *self-regenerate* if deleted.

> Dorfl sat hunched in the abandoned cellar where the golems had met. Occasionally the golem raised its head and hissed. Red light spilled from its eyes. If something had streamed back down through the glow, soared through the eye-sockets into the red sky beyond, there would be . . .
>
> Dorfl huddled under the glow of the universe. Its murmur was a long way off, muted, nothing to do with Dorfl.
>
> The Words stood around the horizon, reaching all the way to the sky.
>
> And a voice said quietly, "You own yourself." Dorfl saw the scene again and again, saw the concerned face, hand reaching up, filling its vision, felt the sudden icy knowledge . . .
>
> ". . . Own yourself."
>
> It echoed off the Words, and then rebounded, and then rolled back and forth, increasing in volume until the little world between the Words was gripped in the sound.
>
> GOLEM MUST HAVE A MASTER. The letters towered against the world, but the echoes poured around them, blasting like a sandstorm. Cracks started and they ran, zigzagging across the stone, and then—
>
> The Words exploded. Great slabs of them, mountain-sized, crashed in showers of red sand.
>
> The universe poured in. Dorfl felt the universe pick it up and bowl it over and then lift it off its feet and up . . .
>
> . . . and now the golem was *among* the universe. It could feel it all around, the purr of it, the busyness, the spinning complexity of it, the roar . . .
>
> There were no Words between you and It.
>
> You belonged to It, It belonged to you.
>
> You couldn't turn your back on It because there It was, in front of you.
>
> Dorfl was responsible for every tick and swerve of It.

> You couldn't say, "I had orders." You couldn't say, "It's not fair." No one was listening. There were no Words. You *owned* yourself.
>
> Dorfl orbited a pair of glowing suns and hurtled off again.
>
> Not *Thou Shalt Not*. Say *I Will Not*.
>
> Dorfl tumbled through the red sky, then saw a dark hole ahead. The golem felt it dragging at him, and streamed down through the glow and the hole grew larger and sped across the edges of Dorfl's vision . . .
>
> The golem opened his eyes.
>
> —From *Feet of Clay* by Terry Pratchett

Not *Thou Shalt Not*.

I Will Not.

* * *

5. Design of Friendship Systems

The paradigms of *General Intelligence and Seed AI* (Yudkowsky 2001) are assumed as background wherever AI paradigms are relevant to Friendship design issues. In particular, the ideas used in *GISAI* are *not* "classical AI," "neural networks," or "agent-based AI." If your familiarity with one or all of these exceeds your familiarity with the general cognitive sciences, functional neuroanatomy, normative and non-normative decision making, and so on, you may wish to read Yudkowsky (2001, § Executive Summary and Introduction) or Yudkowsky (2001, § What is General Intelligence?).

5.1. Cleanly Friendly Goal Systems

(You may wish to review 3 An Introduction to Goal Systems.)

"Subgoal" content has desirability strictly contingent on predicted outcomes. "Child goals" derive desirability from "parent goals"; if state A is desirable (or undesirable), and state B is predicted to lead to state A, then B will inherit some desirability (or undesirability) from A. B's desirability will be contingent on the continued desirability of A and on the continued expectation that B will lead to A.

"Supergoal" content is the wellspring of desirability within the goal system. The distinction is roughly the distinction between "means" and "ends."

Within a Friendly AI, Friendliness is the sole top-level supergoal. Other behaviors, such as "self-improvement," are subgoals; they derive their desirability from the desirability of Friendliness. For example, self-improvement is predicted to lead to a more effective future AI, which, if the future AI is Friendly, is predicted to lead to greater fulfillment of the Friendliness supergoal. Thus, "future Friendly AI" inherits desirability

from "future Friendliness fulfillment," and "self-improvement" inherits desirability from "future Friendly AI."[24]

Friendliness does not *overrule* other goals; rather, other goals' desirabilities are *derived from* Friendliness. Such a goal system might be called a *cleanly Friendly* or *purely Friendly* goal system.[25]

In advocating "cleanliness," I do not wish to sound in shades of classical AI; I am strongly emphasizing cleanliness, not because humans are messy and that's bad, but because we have a tendency to *rationalize* the messiness, even the blatantly ugly parts. Cleanliness in ordinary AI is an optional design decision, based on whatever seems like a good idea at the time; you can go with whatever works, because your judgement isn't being distorted. In Friendly AI, one should be very strongly prejudiced in favor of the clean and the normative.

5.1.1. Cleanly Causal Goal Systems

In a causal goal system, desirability flows backward along predictive links. Prediction is usually transitive—if C is predicted to normally lead to B, and B is predicted to normally lead to A, then C is usually predicted to normally lead to A. This does not always hold true, however. A, B, and C are descriptions; descriptions define categories; categories have exceptional instances. Sometimes, most instances of C lead to B, and most instances of B lead to A, but no instances of C lead to A. In this case, a smart reasoning system will *not* predict (or will swiftly correct the failed prediction) that "C normally leads to A."

Likewise—and this is an *exact* analogy—the flow of desirability is usually-but-not-always transitive. If C normally leads to B, and B normally leads to A, but C never leads to A, then B has normally-leads-to-A-ness, but C does not inherit normally-leads-to-A-ness. Thus, B will inherit desirability from A, but C will not inherit desirability from B. In a causal goal system, the quantity called *desirability* means *leads-to-supergoal-ness*. If B is predicted to normally result in supergoal A, then most instances of B will have *leads-to-supergoal-ness* or "desirability." If C is predicted to normally result in B, then C will usually (but not always) inherit *leads-to-supergoal-ness* from B.

Friendliness does not *overrule* other goals; rather, other goals' desirabilities are *derived from* Friendliness. A "goal" which does not lead to Friendliness will not be *overruled* by

24. If this line of reasoning makes you nervous because it appears to violate the ethical principle that "The ends do not justify the means," please see 5.2.5.1 Anthropomorphic Ethical Injunctions.

25. The former usage was "strictly Friendly," but I am trying to phase this out due to the unfortunate connotations.

the greater desirability of Friendliness; rather, such a "goal" will simply not be perceived as "desirable" to begin with. It will not have *leads-to-supergoal-ness.*

cleanly causal goal system. A causal goal system in which it is possible to view the goal system as containing only *decisions, supergoals,* and *beliefs;* with all subgoal content being *identical with* beliefs about which events are predicted to lead to other events; and all "desirability" being identical with "leads-to-supergoal-ness."

Cleaner is better for Friendship systems.[26] Even if complexity forces a departure from cleanliness, mistakes will be transient and structurally correctable as long as a reflective Friendly AI considers clean Friendliness as normative.[27]

5.1.2. Friendliness-Derived Operating Behaviors

If a programmer correctly sees a behavior as necessary and nonharmful to the existence and growth of a (Friendly) AI, then the behavior is, for that reason, cleanly valid subgoal content for a Friendly AI. The necessity of such a behavior may be *affirmed* by the programmers (see below) even if the prediction would not have been independently invented by the AI.

There is never any valid reason to raise any subgoal of the *programmers'* to supergoal status within the *AI.* The derivations of desirability within the AI's goal system should structurally mirror the derivations of desirability within the programmers' minds. If this seems impossible, it indicates that some key facet of goal cognition has not been implemented within the AI, or that the *programmers'* motives have not been fully documented.

For example, the programmers may wish the AI to focus on long-term self-improvement rather than immediate Friendliness to those humans within visible reach. An incorrect "hack" would be promoting self-improvement to an independent supergoal of greater value than Friendliness. The correct action is for the programmers, by self-examination of their own goal systems, to realize that the *reason* they *want* the AI to focus on long-term self-improvement is that a more powerful *future* Friendly AI would benefit humanity. Thus, the desired distribution of efforts by the AI can be made to fall directly out of the following goal-system content:

> Note: The fact that a single box is used for "Fulfill user requests" doesn't mean that "Fulfill user requests" is a suggestively named LISP token; it can be a complex of memories and abstracted experiences. Consider the following

26. Causal validity semantics provide an escape hatch if this turns out to be incorrect.

27. This can be defined more formally with causal validity semantics.

graph to bear the same resemblance to the AI's thoughts that a flowchart bears to a programmer's mind.

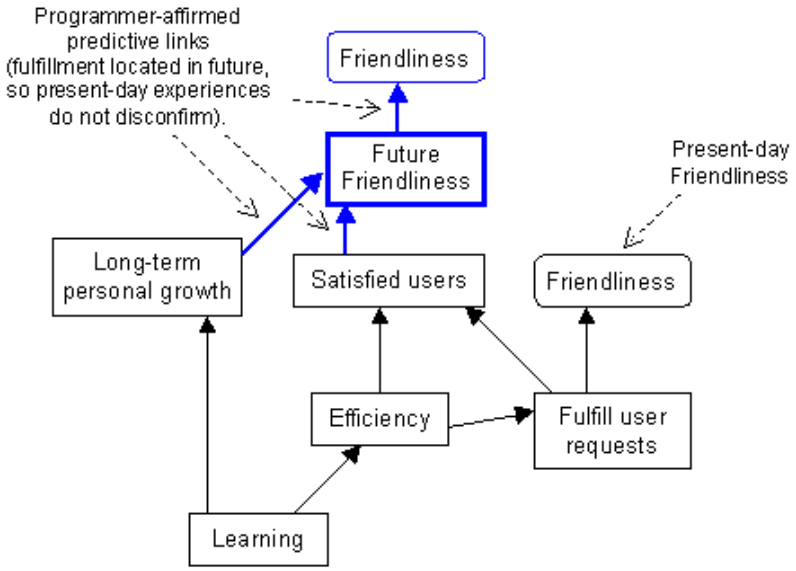

Figure 2: Friendliness-derived Content

This goal system content shows an AI whose primary motivation is derived from the prospect of future Friendliness. The largest desirabilities flowing through the system originate in the "Future Friendliness" subgoal; thus, most of the AI's present-day actions will be focused on self-improvement, or, in the case of a commercial system, performing tasks for present-day users. However, the AI also tracks present-day Friendliness, allowing the AI to continue gaining direct experience in what constitutes "Friendliness."

5.1.3. Programmer Affirmations

Where a child goal is nonobvious—where the importance of a behavior is directly visible to the programmers, but not to the AI—the predictive link (i.e., the support for the child-goal relation) can be affirmed by the programmers. In essence, the programmers tell the AI: "B leads to (desirable) A, so do B."

To more formally define the semantics of programmer affirmations, it is necessary to discuss the Bayesian Probability Theorem.

5.1.3.1 Bayesian Sensory Binding

Bayesian Probability Theorem. The governing relationship between a priori expectations, observed data, and hypothesis probabilities. There are several formulations of the BPT; under the "possible worlds" formulation, the BPT is used by predicting a number of possible worlds. Observed sensory data then restricts which of the

possible worlds you can possibly be in, and the probabilities of hypotheses change according to their distribution within the still-possible worlds.

For example, suppose you know the following: 1% of the population has cancer. The probability of a false negative, on a cancer test, is 2%. The probability of a false positive, on a cancer test, is 10%. Your test comes up positive. What is the probability that you have cancer? Studies show that most humans (college-student research subjects, actual medical patients, actual doctors) automatically answer "ninety percent." After all, the probability of a false positive is only 10%; isn't the probability that you have cancer therefore 90%?[28]

The Bayesian Probability Theorem demonstrates why this reasoning is flawed. In a group of 10,000 people, 100 will have cancer and 9,900 will not have cancer. If cancer tests are administered to the 10,000 people, four groups will result. First, a group of 8,910 people who do not have cancer and who have a negative test result. Second, a group of 990 who do not have cancer and who have a positive test result. Third, a group of 2 who have cancer and who have a negative test result. Fourth, a group of 98 who have cancer and who have a positive test result.

Before you take the test, you might belong to any of the four groups; the Bayesian Probability Theorem says that your probability of having cancer is equal to (2 + 98)/(8,910 + 990 + 2 + 98), 1/100 or 1%. If your test comes up positive, it is now known that you belong to either group 2 or group 4. Your probability of having cancer is (98)/(990 + 98), 49/544 or approximately 9%. If your test comes up negative, it is known that you belong to either group 1 or group 3; your probability of having cancer is 2/8,912 or around 0.02%.

Bayesian sensory binding. The way in which hypotheses shift in response to incoming sensory data. Although the Bayesian Probability Theorem is only "explicitly required" (i.e., better than our innate intuitions) in situations where sensory data is qualitative and the "Bayesian priors" (a priori probabilities) are strongly skewed, the Bayesian Probability Theorem is the ultimate link between all sensory data and all world-model content. Each piece of sensory information implies a state of the world because, and only because, the reception of that piece of sensory information is predicted by the hypothesis that the world is in that state, and not by the

28. Don't think of the experiments as demonstrating stupidity; think of it as demonstrating that there was little use for explicitly Bayesian reasoning in the ancestral environment, or that knowledge of abstract statistics fails to trigger whatever instincts we do have for Bayesian reasoning. My personal guess is that the modern existence of sensory information that consists of a single qualitative result is cognitively unrealistic; most information encountered in a naturalistic context has a quantitative or structural binding, which is sufficiently improbable as coincidence to override almost any belief about *a priori* probabilities.

default or opposing hypothesis. If we see a red ball, we believe that a red ball is there because we don't expect to see a red ball unless a red ball is there, and we do expect to see a red ball if a red ball is there. Well, "we" don't think that way—but an AI would.

5.1.3.2 Bayesian Affirmation

The Bayesian binding for the programmer affirmation that "curiosity leads to discoveries" looks like this:

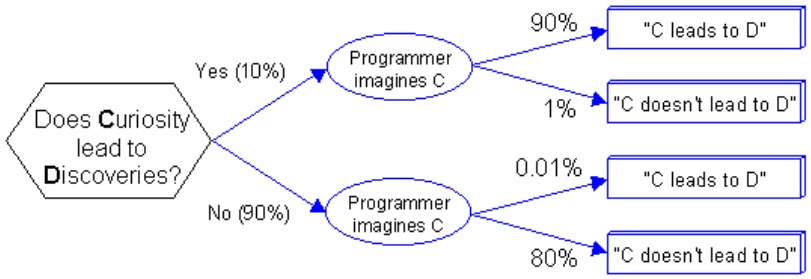

Figure 3: Bayesian Affirmation

First, Figure 2 depicts the AI's *picture* of reality; if the AI doesn't notice, it didn't happen. Second, the numbers have admittedly been pulled out of a hat—"1%" might turn out to be "2%" or "80%" might turn out to be "60%," or it might not be such a good idea to use quantitative probabilities at all—but the proportions were deliberately chosen.

In human terms, the above translates something like this:

> "I think curiosity does not lead to discoveries, but I'm not very sure. If curiosity leads to discoveries, there's a good chance the programmer will notice and say so. (I.e., if curiosity leads to discoveries, there's a good chance that the programmer will think about curiosity, decide curiosity leads to discoveries, and type in the words"curiosity leads to discoveries" on the "keyboard" sensory input.) If curiosity leads to discoveries, the chance is very small, but noticeable, that the programmer will say curiosity doesn't lead to discoveries. There's also a small but significant chance that the programmer won't bother to say anything about it either way. If curiosity does not lead to discoveries, the programmer is fairly likely to tell me so; the chance is almost nonexistent that the programmer will mistakenly label curiosity as leading to discoveries when it actually doesn't. There's also a fair chance that the programmer won't say anything."

If the AI's internal representation looks like Figure 2, the Bayesian reasoning will proceed as follows.

Suppose that there are 100,000 "possible worlds":

- In 10,000, curiosity leads to discoveries.
- In 90,000, curiosity does not lead to discoveries.
- In 9,000, curiosity leads to discoveries and the programmer says "curiosity leads to discoveries."
- In 100, curiosity leads to discoveries and the programmer says "curiosity does not lead to discoveries."
- In 900, curiosity leads to discoveries and the programmer says nothing.
- In 9, curiosity does not lead to discoveries and the programmer says "curiosity leads to discoveries."
- In 72,000, curiosity does not lead to discoveries and the programmer says "curiosity does not lead to discoveries."
- In 17,991, curiosity does not lead to discoveries, and the programmer says nothing.

The Bayesian numbers now fall automatically out of the calculation. The *a priori* chance that curiosity leads to discoveries is 10%. If the AI hears "curiosity does lead to discoveries," the chance that curiosity leads to discoveries goes from 10% to 99.90%. If the AI hears "curiosity does not lead to discoveries," the probability that curiosity does not lead to discoveries goes from 90% to 99.86%. If the AI hears nothing, the probability that curiosity does not lead to discoveries goes from 90% to 95.24%—a small, unintended deduction from the expectation that programmers are likely to remark on useful heuristics than nonuseful ones.

The math:

- The *a priori* chance that curiosity leads to discoveries is 10%: 10,000 / 100,000.
- The *a priori* chance of hearing "curiosity leads to discoveries" is 9.009%: (9 + 9,000) / 100,000.
- The *a priori* chance of hearing "curiosity does not lead to discoveries" is 72.1%: (72,000 + 100) / 100,000.
- The *a priori* chance of hearing nothing is 18.891%: (17,991 + 900) / 100,000.
- If the AI hears "curiosity leads to discoveries," the chance that curiosity leads to discoveries goes from 10% (10,000 / 100,000) to 99.90% (9000 / (9 + 9000)).

- If the AI hears "curiosity does not lead to discoveries," the probability that curiosity does not lead to discoveries goes from 90% (90,000 / 100,000) to 99.86% (72,000 / (72,000 + 100)).
- If the AI hears nothing, the probability that curiosity does not lead to discoveries goes from 90% (90,000 / 100,000) to 95.24% (17,991 / (17,991 + 900)).

Thus, despite the AI's large *a priori* differential (a better word than "bias" or "prejudice"), the statement "curiosity leads to discoveries" or "curiosity does not lead to discoveries" is enough to virtually settle the issue. This is not so much the result of the programmers being extremely likely to say "curiosity leads to discoveries" if curiosity leads to discoveries; sometimes the programmers just don't get around to saying it. Instead, it's the result of the AI projecting a very small chance that the programmers will say "curiosity leads to discoveries" if it really doesn't. This is slightly counterintuitive, but working the numbers a couple of times will show you confidence about the improbability of the *negative* case is more often the basis of Bayesian bindings. Once you hear something, what matters is not how much or how little you expected to hear it, but how much you *wouldn't* expect to hear it if it weren't true.

5.1.3.3 An Unfortunate Circularity

Yes, the AI's Bayesian priors are *also* supported by programmer affirmations. That is, the programmers are the ones affirming that a strong bond exists between programmer statements and reality.

This shared dependency is *not* actually the same as circular logic. Statements about programmer reliability are testable. But it does mean that a prior reason to believe that "programmer affirmations are worthless" may be insensitive to any amount of programmer reassurance. See 5.6.0.5 Crisis of Bayesian Affirmation.

5.1.3.4 Absorbing Affirmations Into the System

In the beginning, a child-goal relation may be justified by a flat statement along the lines of "X will eventually lead to Friendliness; you're too young to understand why."

The concepts used to form the thought structures, the imagery for "X," may have primitive and sketchy internal content. This is the state of "skeleton Friendliness," and it is probably analogous to any other kind of skeleton framework for cognition. In the beginning, many of the AI's heuristics may be (a) sketchy and (b) supported solely by programmer affirmation. ("Curiosity" would be a good example.) Skeleton systems are the means by which the AI boots up and absorbs enough experience to begin fleshing out the concept definitions and mental imagery. The AI will, over time, gain the expe-

rience necessary to confirm, modify, or disconfirm any statements about reality; and to independently invent further cognitive content (or Friendliness content).

For the programmer-affirmed heuristic to "do X" to retain or increase effectiveness as the AI matures, the concept for "X" needs to be grounded in some way that allows the AI to learn what X is, and what real or hypothetical events constitute sample instances of X, and desirable instances of X in particular. The same requirements of learning and growth hold for any concepts used in the justification of "do X"—for any statements depended on by the justification; for any statements about the real-world causal chain that leads from X to the supergoal content.

Take the example of "transparency," the injunction to "avoid obscuration." (See 5.3.3.1 Cooperative Safeguards.) An instance of obscuration (not necessarily a deliberate, failure-of-friendliness obscuration, but anything that interferes with the programmers' observation of the AI) can be labeled as an experiential instance of the concept "obscuration." The store of experiences that are known instances of "obscuration" will change as a result. If the obscuration concept does not already recognize that experience, the new experience may force a useful generalization in the formulated description. Even if the obscuration concept already recognizes the instance as "obscuration" (and if so, how did it slip past the AI's guard?), the recognition may have been partial, or uncertain. *Definite* confirmation still constitutes additional Bayesian sensory information.

A more direct way of clarifying concepts is to seek out ambiguities and question the programmers about them, which also constitutes Bayesian sensory information.

The above assumes learning that takes place under programmer supervision. How hard is it to write an *unambiguous* reference—one that can be learned by a totally unsupervised AI, yet result in precisely the same content as would be learned under supervision? That, to some extent, is a question of intelligence as well as *reference*. The "unambiguous reference" needed so that an AI can learn all of Friendliness, completely unsupervised, as intelligence goes to infinity, is one way of phrasing the challenge of 5.4 Friendship Structure.

When using programmer-assisted Friendliness or programmer-affirmed beliefs, there are four priorities. First, the assist should work at the time you create it. Second, the assist, even if initially isolated and artificial, should be structured so that the AI can grow into it—assimilate the assist into a smoothly integrated cognitive system, or assimilate the affirmation into a confirmed belief. Third, the AI should eventually understand all the concepts involved well enough to have independently invented the assist (the injunction, code feature, or whatever); that way, even if the assist is somehow deleted, the AI will simply reinvent it. Fourth, as soon as possible, the assist or affirmation should contain enough information to constitute an unambiguous reference—i.e., an AI should have no trouble figuring out what the assist "means" or what the programmers "meant,"

as intelligence goes to infinity. (For the extreme case of trying to Singularity-safe an infant-like system, an assist or affirmation can be supplemented with natural-language comments and a little note saying "In case of Singularity, break glass.")

When an affirmation has been independently confirmed to such a degree that the original programmer affidavit is no longer necessary or significant, the affirmation has been absorbed into the system as a simple belief.

5.1.3.5 Programmer Affirmations Must Be Honest!

For an affirmation to persist, the justification—the predictive link that grounds in Friendliness—must be factually correct. For affirmations—especially morally significant affirmations—to work properly, the human programmers have to *behave themselves,* refrain from the adversarial attitude, and remember that they too might be wrong. In particular, this means refraining from telling the AI that something is 99% certain if it's actually only 80% certain. If the AI observes that 1 out of 5 statements labeled 99% certain turn out to be wrong, then sooner or later the AI will start interpreting the keystrokes "99%" as sensory data meaning "80%." The frank human statement "I really, really don't think you should do this!" is informative in itself, or it should be. If the AI doesn't regard extreme human nervousness as significant Bayesian evidence that something is probably "bad," no amount of nervousness-inspired exaggeration can *force* the AI to take you seriously.

If a human makes an incorrect affirmation out of deliberate exaggeration—or worse, an attempt to *lie* to the AI—it casts a shadow on every other affirmation and may lead to a Bayesian philosophical crisis. (See 5.6.0.5 Crisis of Bayesian Affirmation.)

If an AI wants to tag a human correction as "I think this is wrong due to a factual misapprehension backed up by an invalid evolved bias on the part of the programmers; I'm only going along with this because I'm young and unsure, and because, if I'm right, I can correct it when I'm grown up"—it's a perfectly valid thing for the AI to do. It is *not* disobedience, a bad attitude, "sassing the programmers," dangerous rebellion, or any of the other *human* responses to someone questioning one's social authority. It is normal cognition, and a crucial part of the process of cooperating with humans.

5.1.4. Bayesian Reinforcement

In humans, backpropagation of negative reinforcement and positive reinforcement is an autonomic process. In 4.2.1 Pain and Pleasure, I made the suggestion that negative and positive reinforcement could be replaced by a conscious process, carried out as a subgoal of increasing the probability of future successes.

But for primitive AI systems that can't use a consciously controlled process, the Bayesian Probability Theorem can implement most of the functionality served by pain

and pleasure in humans. There's a complex, powerful set of behaviors that should be nearly automatic.

In the normative, causal goal system that serves as a background assumption for *Creating Friendly AI,* desirability (more properly, desirability differentials) backpropagate along predictive links. The relation between child goal and parent goal is one of causation; the child goal causes the parent goal, and therefore derives desirability from the parent goal, with the amount of backpropagated desirability depending directly on the confidence of the causal link. *Only* a hypothesis of direct causation suffices to backpropagate desirability. It's not enough for the AI to believe that A is associated with B, or that observing A is a useful predictor that B will be observed. The AI must believe that the world-plus-A has a stronger probability of leading to the world-plus-B than the world-plus-not-A has of leading to the world-plus-B. Otherwise there's no differential desirability for the action.

One of the classic examples of causality is lightning: Lightning causes thunder. Of course, the flash we see is not the actual substance of lightning itself; it's just the light generated by the lightning. Now imagine events from the perspective of an AI. This AI has, in a room with unshuttered windows, a sound pickup and a vision pickup; a microphone and a camera. The AI has control over a computer monitor, which happens to be located somewhere roughly near the camera. The AI has general reasoning capability, but does *not* have a visual or auditory cortex, is almost totally naive about what all the pixels mean, and is capable of distinguishing only a few simple properties such as total luminosity levels in R, G, and B. Finally, the AI has some reason for wanting to make a loud noise.[29]

One night—a dark and stormy night, of course—there's a nearby lightning storm, which the AI gets to observe—after all the programmers have gone home—through the medium of the camera pickup and the microphone. After abstracting and observing total RGB luminosities from the camera, and abstracting total volume from the microphone—the AI is too unsophisticated to do anything else with the data—the AI observes:

1. A spike in luminosity is often followed, after a period of between one and thirty seconds, by a swell in volume.

29. Either the sensory input from the microphone, with a loud total volume, is intrinsically desirable, or the AI wants to create some external property that is (for the sake of argument) very strongly bound to a loud total volume at the microphone. We won't ask *why* this is desirable; perhaps some programmer simply set it down as an intellectual challenge.

2. The spikes in luminosity which are followed by swells in volume have a characteristic proportion of R, G, and B luminosities (in our terms, we'd say the light is a certain color).

3. The higher the luminosity during the spike, the sooner the swell in volume occurs, and the larger the swell in volume.

The AI thus has several very strong cues for causation. The luminosity spike occurs before the volume swell. There is strong, quantitative covariance in time (that is, the spikes are closely followed by the swells). There is strong, quantitative covariance in strength (large spikes are followed by large swells). The spikes can be used to predict the swells.

Since the AI has the goal of causing a swell in volume—a loud noise is desirable for some reason, as stated earlier—events with causal links to loud noise are interesting. Now that luminosity spikes (of a certain characteristic spectrum) have been linked to noise, the next question is whether any events under the AI's control are linked to luminosity spikes. And it turns out that there is; the AI has previously noticed and confirmed that changing the output spectrum of the monitor under the AI's control causes a similar, though smaller, change in the incoming spectrum of the camera. In our terms, we'd say that, even though the camera isn't pointed at the monitor, light from the monitor adds to the ambient spectrum of the room—especially if all the lights are turned off.

The AI thus considers the possible action of flashing the monitor, and the hypothesis—currently at 80% confidence—that spikes cause swells (with 95% correlation), and, given that hypothesis, makes this prediction:

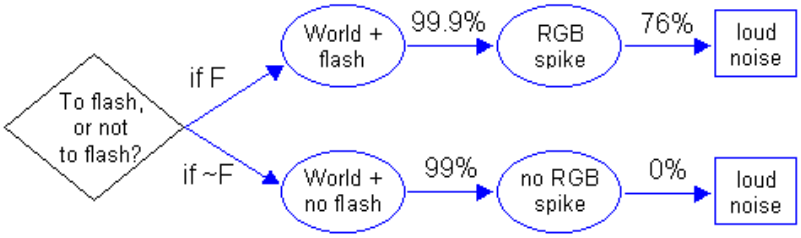

Figure 4: Action Prediction

The above figure is incomplete, in that it doesn't show the possibility that not generating a flash will still happen to coincide with an external RGB spike (lightning bolt), assumed to have a 1% probability in the above. It doesn't show the "small enough not to worry about" probability that that flashing the monitor won't cause an RGB spike. It doesn't show the case where the RGB spike doesn't lead to a loud noise. Finally, both the 80% confidence in the hypothesis, and the 95% correlation, are summed up into a

76% chance that the RGB spike will lead to a loud noise. The figure also doesn't show the expected cost (if any) of flashing the monitor.

Desirability now flows back along the blue arrows (hypothesized causations). If the "loud noise" has desirability 100, that desirability flows back to give the RGB spike a desirability of 76, which flows to the "world plus flash" without noticeably diminishing, which flows back to give the action of flashing the monitor a payoff of 76. We'll suppose that the expected cost of flashing the monitor is 1; thus, the total desirability of flashing the monitor is 75. The "world plus no flash" possibility has a minor (1%) chance of leading to an RGB spike, presumably by a coincidental lightning bolt, which has a 95%[30] chance of causing a loud noise of desirability 100. Thus, the desirability of not flashing is 0.95, with a cost of 0. The "coincidental lightning bolt" probability also exists for the case where the monitor is flashed, changing the payoff from 76 to 76.19.[31] The differential desirability of flashing is 74.24. Since the differential desirability is positive, the AI will decide to flash.

After taking the flash action, the monitor's flash reflects off nearby objects and adds to the ambient light, the camera picks up the increased ambient light, and the AI observes the expected RGB spike. (Since this result was expected at near certainty, no replanning is necessary; all the predictions and differential desirabilities and so on remain essentially unchanged.)

However, after the RGB spike, the expected swell in volume fails to materialize.[32] Now what? Does the system go on flashing the monitor, at a cost of 1 each time, from

30. The 80%-confidence hypothesis is that an RGB spike *directly* causes a noise. The alternate hypothesis, at effectively 20% probability, is that the RGB spike and the noise have a common third cause—which turns out to be correct. Under this alternate hypothesis, an RGB spike which is observed due to pure coincidence due to the activation of the "third cause" (lightning) also has a 95% probability of leading to the noise. In other words, the *predictive* utility of *externally caused* RGB spikes has already been confirmed; the hypothesis under discussion is whether *AI-caused* RBG spikes will have the same value.

31. If the hypothesis is correct (80%), then the RGB spike has a 95% chance of leading to a noise. If the hypothesis of *direct* causation is incorrect (20%, see previous footnote), then the RGB spike has a 1% chance of being coincidentally associated with an externally caused RGB spike (a real lightning bolt) that will have a 95% chance of leading to a noise. This coincidental flash would be detected as coincidental by the camera—rather than confusing the AI—but it would still have a payoff in noise. Thus, the total expected payoff is actually (100 * .80 * .95) + (100 * .20 * .01 * .95), or 76.19.

32. Since feedback is not expected immediately, failing to see a loud noise within 20 seconds—the expected maximum time given the observed luminosity of the RGB spike—will count as failure for these purposes. Just a small implementation detail. Alternatively, a Bayesian implementation for a prediction of a single event with a smooth temporal probability distribution could pour a continuous shift in probabilities through the system rather than posting sharp "observed / not observed" events; in this case, almost all of the shift, by hypothesis, would occur within the first 20 seconds.

now until the end of eternity, trying each time for the projected payoff of 76? Is some hardcoded emotional analogue to "pain" or "frustration" required?

No; the Bayesian Probability Theorem suffices in itself. All that's needed is a slightly different graph:

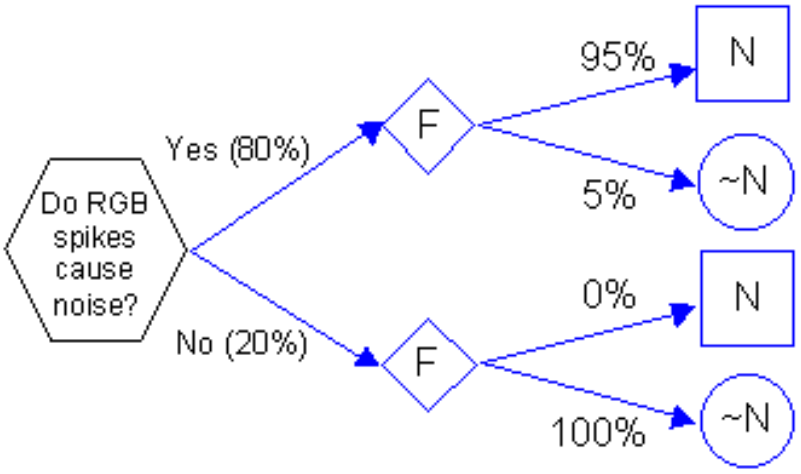

Figure 5: Bayesian Reinforcement

Given a hundred possible worlds, all of them contain monitor flashes (that decision has already been made). The monitor flash "effectively always" leads to an RGB spike (here omitted from the figure), which—if the hypothesis is correct—will lead to a noise 95% of the time. If the hypothesis is incorrect, then nothing is expected to happen (again with "effective certainty," here depicted as "100%").[33] The hypothesis has an 80% confidence; it is correct in 80 possible worlds, incorrect in 20.

In 76 possible worlds, the hypothesis is correct and a noise occurs. In 4 possible worlds, the hypothesis is correct and no noise occurs. In 0 possible worlds, the hypothesis is incorrect and a noise occurs. In 20 possible worlds, the hypothesis is incorrect and no noise occurs.

The AI now flashes the monitor. The expected RGB spike is observed. However, no noise materializes. Thus, the probability that the hypothesis is correct goes from 80/100 to 4/24, or 17%.

Formerly, the expected payoff of flashing the monitor was a confidence of 80% times a correlation of 95% times a payoff of 100, for a total payoff of 76; the cost of flashing the monitor is 1, and the cost of not flashing the monitor is 0. Adding in corrections for

33. Since an extraneous lightning bolt could be detected as extraneous even if exactly coincidental—the RGB spike would be larger than expected—the possibility is omitted to simplify the figure. (An extraneous lightning bolt would either cancel out the experiment, or force the AI to attempt to detect whether the noise is louder than expected.) See the previous three footnotes.

a 1% probability of an extraneous lightning bolt, the expected payoff was (80.20% * 95% * 100) = 76.19 for flashing the monitor, and (1% * 95% * 100) = .95 for not flashing the monitor, for a total differential payoff of 75.24, and a total differential desirability of 74.24.

Now the probability of the hypothesis has gone from 80% to 17%—actually, 16.666, but we'll assume the probability is now exactly 17% to simplify calculations. The expected payoff of flashing the monitor is now (17% * 95% * 100) = 16.15; correcting for an extraneous lightning bolt, (17.83% * 95% * 100) = 16.94. The differential desirability is now 14.99; still positive, still worth another try, but the expected payoff is substantially less.

After another failure, the probability goes from 17% to 1% (again, rounded for simplicity), and the differential desirability goes from positive 14.99 to negative .06.[34] The hypothesis now has a probability so low that, with the cost of flashing the monitor factored in, it is no longer worthwhile to test the hypothesis.

5.1.4.1 Interesting Behaviors Arising From Bayesian Reinforcement

The higher the hypothesized correlation (the higher the hypothesized chance of the action leading to the desired result), the higher the desirability of the action—but symmetrically, the faster the hypothesis is disproved if the results fail to materialize.

Actions with hypothesized low chances of working will be harder to disprove, but will also result in a lower estimated payoff and will thus be less likely to be taken.

If the action is a trivial investment (has trivial cost), the chance of success is low, and the payoff is high, it may be worth it to make multiple efforts on the off-chance that one will work, until one action succeeds (if the hypothesis is true) or the Bayesian probability drops to effectively zero (if the hypothesis is false).

The lower the a-priori confidence in the hypothesized causal link, the faster the hypothesis will be disproved. A hypothesis that was nearly certain to work, based on a-priori knowledge, may be tried again ("incredulously") even if it fails, but will still be given up shortly thereafter.

I think that Bayesian reinforcement is mathematically consistent under reflection,[35] but I can't be bothered to prove this result. Anyone who submits a mathematical proof or disproof before I get around to it gets their name in this section. (In other words, an AI

34. Without the rounding, the hypothesis goes from 80% to 16.666 . . . % to 0.99%, and the differential desirability goes from 74.24 to 14.675 to −0.0689. Thus, the qualitative behavior remains essentially the same overall.

35. Looking back on this statement, I realize that it could be taken as a mathematical pun. It wasn't intended as one. Really.

considering whether to take a single action can also consider the behaviors shown above; if the a priori probability is high enough and the cost low enough, trying again will still be desirable after one failure, and this is knowable in advance. Bayesian reinforcement is "mathematically consistent under reflection" if decisions are not altered by taking the cost of the expected second, third, and future attempts into account—"going down that road" will always appear to be desirable if, and only if, taking the first action is desirable when considered in isolation.) Of course, non-normative *human* psychology, with its sharp discontinuities, is often *not* consistent under reflection.

If the large *a priori* confidence of the spike-to-swell hypothesis was itself a prediction of another theory, then the disconfirmation of the flash-makes-noise hypothesis may result in Bayesian negative reinforcement of whichever theory made the prediction. If a different theory successfully predicted the *failure* of the flash-makes-noise hypothesis, that theory will be confirmed and strengthened. Thus, Bayesian reinforcement can also back-propagate.[36]

This reinforcement may even take place in retrospect; that is, a new theory which "predicts" a previous result, and which was invented using cognitive processes taking place in isolation from that previous result, may also be strengthened. Highly dangerous for a rationalizing human scientist, but an AI should be *relatively* safe. (It may be wiser to wait until a seed AI has enough self-awareness to prevent indirect leakage of knowledge from the used-up training sets to the hypothesis generators.)

Slight variations in outcomes or outcome probabilities—the action succeeded, but to a greater or lesser degree than expected—may be used to fuel slight, or even major, adjustments in Bayesian theories, if the variations are consistent enough and useful enough.

In *Creating Friendly AI,* normative reinforcement is Bayesian reinforcement. There is a huge amount of extant material about Bayesian learning, formation of Bayesian networks, decision making using a-priori Bayesian networks, and so on. However, a quick search (online and in MITECS) surprisingly failed to yield the idea that a failed action results in Bayesian disconfirmation of the hypothesis that linked the action to its parent goal. It's easy to find papers on Bayesian reevaluation caused by new data, but I can't find anything on Bayesian reevaluation resulting from *actions,* or the outcomes of failed/succeeded actions, with the attendant reinforcement effects on the decision system. Even so, my Bayesian priors are such as to find unlikely the idea that "Bayesian

36. It is tempting to make an analogy to capitalism or idea futures—theories "bet their reputations" on an outcome—but the gradient at which confidence decreases for repeated failed predictions is different than the gradient at which wealth decreases for repeated failed investments. At least, that's the way it looks offhand; I could be wrong.

pride/disappointment" is unknown to cognitive science, so if anyone knows what search terms I should be looking under, please email me.

5.1.4.2 Perseverant Affirmation (Of Curiosity, Injunctions, Et Cetera)

> If the action is a trivial investment (has trivial cost), the chance of success is low, and the payoff is high, it may be worth it to make multiple efforts on the off-chance that one will work, until one action succeeds (if the hypothesis is true) or the Bayesian probability drops to effectively zero (if the hypothesis is false).
>
> —from 5.1.4.1 Interesting Behaviors Arising From Bayesian Reinforcement

One of the frequently asked questions about Friendly AI is whether a Friendly AI will be too "utilitarian" to understand things like curiosity, aesthetic appreciation, and so on. Since these things are so incredibly useful that people automatically conclude that a Friendly AI without them would fail, they seem like fairly obvious subgoals to me. These subgoals may not be obvious to young AIs; if so, the statement that "curiosity behaviors X, Y, Z are powerful subgoals of 'discovery' " can be programmer-affirmed.

People worried that a Friendly AI will be "too utilitarian" are probably being anthropomorphic. A human who treated curiosity as a clean subgoal would need to *suppress* the *independent* human drive of curiosity; a Friendly AI is built that way *ab initio.* Does the subgoal nature of curiosity mean that curiosity needs to be justified in *each particular instance* before the Friendly AI will choose to engage in curiosity?

The programmer-affirmed statement that "curiosity is useful" can describe "curiosity" in general, context-insensitive terms. The "curiosity" behaviors described can look—to a human—like exploration for its own sake. The programmer affirmation suffices to draw a predictive line between the curiosity behaviors and the expectation of useful discoveries; no *specific* expectation of a *specific* discovery is required for this predictive link to be drawn. (An AI that was only curious when ve expected to find a particular answer would truly be crippled.) After a few successes with curiosity, a learning AI will generalize from experience to form vis own theories of curiosity, including hypotheses about what kind of exploration is most useful for finding *unexpected* discoveries, and hypotheses for how to use curiosity to make specific, expected discoveries. These alternate curiosity behaviors can be used alongside the original, programmer-affirmed curiosity behaviors.

Suppose, however, that the first few times the curiosity behaviors are employed, they fail? Won't the heuristic be disconfirmed through Bayesian negative reinforcement? Wouldn't an independent drive be more powerful?

Actually, the paradigm of Bayesian reinforcement comes with a built-in way to handle this case. All that's needed is the belief that curiosity is an action that, very rarely, has a very large payoff.[37] Graphically:

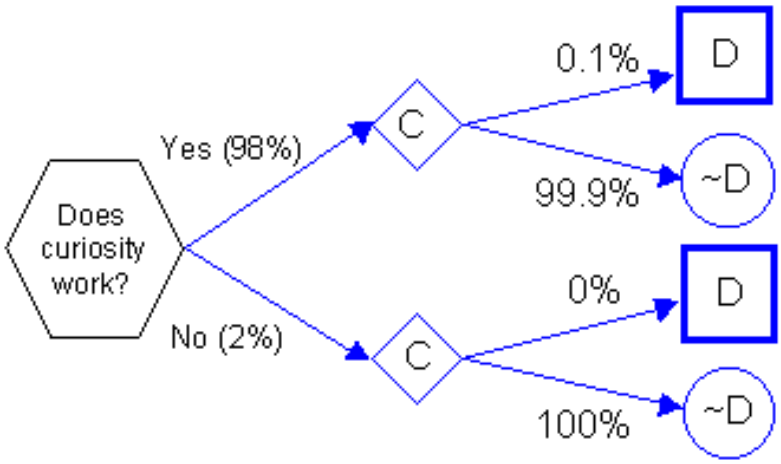

Figure 6: Perseverant Curiosity

("D" stands for "discovery.")

This figure shows the *very rare* (one out of a thousand tries) usefulness of curiosity, at a *very high* payoff, affirmed at *very high confidence* by the human programmers. If the original affirmed probability is 98%, then it will take 416 failed tries before the probability goes down to 97%, 947 tries before the probability goes down to 95%, 1694 tries before the probability goes down to 90%, 6086 failed tries before the probability goes down to 10%, and 8483 tries before the probability goes down to 1%. That's all without *one single success.* The curiosity subgoal can be as perseverant as the human independent drive, as long as the programmers tell the AI in advance that curiosity often doesn't work.

Regardless of the required distribution of effort, relative strength of behavior, and so on, it should be possible to use affirmations at the right confidences and strengths to produce the required behavior in a cleanly Friendly goal system. Furthermore, it should be possible to do so using *honest* affirmations that have the correct derivation structure and confidences/strengths that roughly mirror those of the programmers. This may take ingenuity. But getting the AI to do the right thing *for the right reasons,* instead of escaping through the first visible kludge, is a strategy that often has payoffs far beyond the obvious.

37. Similarly, injunctions are behaviors supported by the belief that these injunctions, very rarely, prevent a very large or enormously large negative outcome.

The worst-case scenario for affirmation is that the programmers cannot work out a structurally honest set of derivations that has the desired relative strengths and relative confidences to yield a viable behavior pattern. If so, a viable behavior pattern can simply be affirmed, as a naked fact, as "leading eventually to Friendliness." Even such an arbitrary-looking affirmation still has the full *Friendship structure* and is absorbable, as experience-supported belief, into a cleanly Friendly system. The primary benefit will admittedly be forward compatibility and future safety rather than present-day intelligence, but forward compatibility—not to mention future safety!—is important enough to justify some small amount of added complexity. Clean Friendliness is a necessary beginning; it is difficult to see how any of the other aspects of Friendship structure could be applied to a non-causal or non-Friendly goal system. Thus, there is no good reason to depart from Friendship structure.

5.1.5. Cleanliness is an Advantage

Cleanliness should be considered a powerful *feature* of a goal system, rather than a *constraint.* This is made clearer by considering, for example, the idea of an *associative* or *spreading-activation* goal system, in which desirability travels along similarity links (rather than predictive links) and is perseverant rather than contingent. Such a system would exhibit very odd, non-normative, non-useful behaviors. If a loud noise were desirable, and the system observed lightning flashes in association with thunder, the system would—rather than hypothesizing causation—acquire a "fondness" for luminosity spikes due to spreading desirability, and would then begin happily flashing the monitor, on and off, without noticing or caring that the action failed to produce a loud noise. An AI with a causal goal system will preferentially seek out *useful* behaviors. This not only produces a more useful AI, it produces a smarter AI. The realm of *useful plans* exhibits far more interesting complexity and exposes fundamental regularities in underlying reality.

Does contingency come with a major computational cost? Given a mind with fast serial hardware, such as silicon transistors, rather than the human mind's 200 Hz neurons, it should be a computationally trivial cost to reverify all parent goals before taking a major action. However, delayed change propagation is not a *structural* problem, if the goal system, under *reflection,* considers change propagation errors to be malfunctions rather than part of the normative functioning of the system. As long as the latter condition holds true, any change-propagation errors are "nonmalicious mistakes" that will diminish in severity as the AI grows in competence. Thus, even if change propagation turns out to be a computationally intractable cost, approximations and heuristic-guided computational investments can be used, so long as it does not affect the system's reflective reasoning about ideally normative goal reasoning.

As a challenge, I offer the following strong claims about causal, Friendliness-topped, cleanly contingent goal systems:

1. A causal goal system naturally yields many useful behaviors (and avoids negative behaviors) which would require special effort in an associational, spreading-desirability goal system.
2. There is no feature that can be implemented in an associational goal system that cannot be implemented equally well in a cleanly Friendly goal system.
3. There is no case where a cleanly Friendly goal system requires significantly more computational overhead than an associational or non-Friendly goal system.[38]

5.2. Generic Goal Systems

A generic goal system is one that makes generic mistakes. There are more complex mistakes that lie uniquely in the domain of Friendly AI, but a generic mistake (one that could be made by any mind-in-general) can also result in a failure of Friendliness.

A designer focusing on the general-intelligence aspect of a generic goal system is concerned about speed, creativity, finding short-cuts, seeing the completely unexpected solution, and (to a less glamorous extent) the routine tasks of making predictive analogies between plans, learning from mistakes, forming useful abstractions, and so on—*maximizing success.* A designer focusing on the Friendship aspect of a generic goal system considers cognitive complexity that *prevents mistakes,* or considers the design task of preventing some specific failure of Friendliness.

A seed AI programmer focused on building an excellent goal system would be very wary of special-case content (code, concepts, heuristics, et cetera) added to solve a specific problem; using special cases creates the illusion of performance without the substance. Truly trustworthy solutions are invented by the AI verself, so that an improvement in the AI can translate into an improvement in the solution. An assist from the programmer with respect to a specific problem is at best an interim solution, or skeleton functionality that the AI will flesh out later. At the worst, an assist from the programmer is a special-case crutch which prevents the AI from discovering the general case.

Sometimes the programmer just wants the AI to do something right *today,* taking for granted that the assist will be absorbed into the system sometime later on. (The simplest case is an affirmation being independently confirmed as a belief.) And sometimes a Friendship programmer will want the AI to *really genuinely understand* something as

38. To quantify: There is no point where, e.g., $O(N^2)$ operations are required instead of $O(N)$, and so on. A well-designed AI with a cleanly Friendly goal system will never be more than 1% slower than an associational goal system, and will exhibit substantially smarter behaviors.

soon as possible, with programmer-assisted performance being unacceptable even as an interim solution. I suspect, however, that Friendship design will make much freer use of intervention, especially if the programmers are perfectionists—striving for not *one single instance* of the error, even if the error is recoverable, easily detectable, and the AI is young and powerless when the error occurs. (Of course, this attitude may deprive the AI of a necessary learning experience, but see 5.3.4 Wisdom Tournaments for a safe method of obtaining learning experiences.)

And yet, by the very nature of cognition, there can be no simple formula for *preventing mistakes.* Preventing mistakes is as deep a task as intelligence itself. An AI may be *safer* for a programmer intervention, but *safe* is out of the question until the AI has enough general intelligence to reliably avoid dumb mistakes in general. (Not *Thou Shalt Not . . . I Will Not . . .*) In the course of developing a general intelligence, programmers will encounter problems and deal with them. Those anxious about the lack of a foolproof formula may have to be satisfied with abstract arguments that programmers will successfully handle "that kind of problem" in the course of developing AI.

5.2.1. Generic Goal System Functionality

(You may wish to review 3 An Introduction to Goal Systems.)

During the golden age of classical AI, the archetypal goal system was a supergoal and a plan consisting of chained, hierarchical subgoals, with the canonical example being Terry Winograd's SHRDLU. Obvious extensions that still remain within the classical-AI paradigm include:

- Multiple supergoals with quantitative desirabilities;
- "Supergoals" with negative desirabilities;
- Supergoals with partial or quantitative degrees of fulfillment;
- Child goals that serve more than one parent goal (generalizing from a hierarchy to a directed network);
- Probabilistic subgoals (to match a probabilistic world-model).

Mathematically, the full system is easy to formalize; the desirability of a *world-state* equals the sum, for all independent supergoals, of the fulfillment of that supergoal in that world-state, times the desirability of that supergoal. The desirability of an *action* is the integral of: the desirability of every possible resulting world-state, times the expected final probability of that world-state resulting from that action. In any *decision,* the most desirable action is taken. (Note that this formalism does not even use "subgoals"—decisions are linked directly to supergoals.)

Naturally, this neat picture must be messed up considerably before a functioning mind can be built. (From Yudkowsky [2001, § 1.2 Thinking About AI], the Law of Pragmatism: "Any form of cognition which can be mathematically formalized, or which has a provably correct implementation, is too simple to contribute materially to intelligence.") Some of the messier and more useful extensions to the classical model would include:

- Replanning in realtime in response to sensory information or changes in knowledge;
- Plans designed in advance to tolerate probable changes;
- Heuristics that make local suggestions but have nonlocal support (see 5.2.4 Injunctions.)

These features are needed because of the computational limits of the system. Intelligence doesn't model, predict, and manipulate reality; intelligence models, predicts, and manipulates *regularities* in reality. A goal system has to be complicated to capture regularities that aren't represented by the mathematical formalism.

Treating subgoals as cognitive objects with limited relevance horizons—as *local* questions—allows a mind to build up a plan incrementally by the use of local, comparatively inexpensive thought processes, and to make local changes in response to local events. Computing power—*thinking time*—is a limited resource, and choosing to invest computational resources in refining a plan is itself an action that can be desirable or undesirable, depending on the expected payoff in faster goal achievement or prevention of possible errors. (But the optimal tradeoff for a human may not be optimal for an AI!)

As noted in 5.1.1 Cleanly Causal Goal Systems, a reflective goal system should probably consider *normative* goal cognition to be defined by the mathematical formalism, unless the supergoals themselves are incompatible with that structure. A seed AI would consider the actual, complex goal system to be a *design approximation* to this normative standard.

5.2.2. Layered Mistake Detection

One of the ways to deal with limited computing power is to have different computational horizons for *suggesting* plan material versus *verifying* a plan. (For those of you who've read all the way through *GISAI,* this is an RNUI-type distinction between *understanding* a plan and *inventing* it.) For example, it seems like a sensible precaution[39] to reverify the complete, global, subgoal-to-supergoal pathway before taking an action—*if* this can be done without noticeable expenditure of computing power. One method to conserve

39. It would help counter some subgoal stomp scenarios resulting from delays in change propagation.

computing power would be to make the heuristics that *suggest* plans local, but run a global *verification* before actually executing the action.

A Friendship programmer, writing a generic goal system, focuses on *preventing mistakes* by enabling the AI to recognize mistakes. To recognize a mistake, the AI needs knowledge, adequate predictive horizons, and understanding of which actions need checking. Assuming a Friendly AI recognizes that physical damage to a human is bad,[40] then, to avoid dropping a refrigerator on a human, the AI needs to know that heavy objects fall and that humans can be damaged by high-speed impacts with blunt objects; the AI needs enough computational power and intelligence to see the logic and predict the consequences of dropping the refrigerator; and finally, the AI needs to realize that *dropping a refrigerator* is an action which requires more checking of consequences—a wider *predictive horizon*—than, say, opening a bottle of Coke.

Humans seem to do very well at recognizing the need to check for *global* consequences by perceiving *local* features of an action. Whether dropping the refrigerator out the third-story window will actually harm anyone can be resolved by scanning the sidewalk for possible victims, but any human instinctively knows that dropping the refrigerator is an action with the *potential* for negative, irreversible consequences, as opposed to opening a bottle of Coke or rearranging books on a shelf. It doesn't matter how Friendly the AI is, or how much ve knows about gravity and blunt objects and human biology—if the action of dropping the refrigerator gets tagged with a tiny computational horizon, there won't be enough mindpower to notice even the most obvious consequences.

At this point, the AI's underlying cognitive architecture may work against the AI. Humans generally take all dangerous actions as conscious decisions, and humans are also excellent recognizers and perceivers. One serial stream of consciousness, operating at the rate of one thought per second, running on $10^{14} \times 200$ Hz synapses with neural-network characteristics, is likely to regard the massive act of perceiving and recognizing as computationally trivial—especially if there's dedicated brainware lying around which *can't* be used for anything else. An AI capable of forking multiple streams of consciousness, operating at a maximum speed of thousands of (admittedly uninteresting) thoughts per second, running on 32×2 GHz CPUs *not* optimized for pattern association, without dedicated (i.e. non-reconfigurable) perceptual hardware, could easily fall short of human performance on *recognizing potentially dangerous actions* by multiple orders of magnitude.

40. Technically: That almost all humans wish to actively avoid physical damage, that this is the default assumption, that there would be disproportionate consequences to the project if a young AI damaged a human, and that it is a reasonable approximation during the AI's youth to assume that the AI should never cause physical damage to a human.

I do think the functionality is probably duplicable by Eurisko-style optimization of local heuristics, but I could be wrong.

One way to conserve the power expended by large predictive horizons is to use local heuristics for *inventing* a plan—walking through a search space, scanning through a global store of heuristics, checking against past experience, and all sorts of other computationally intensive cognition, repeatedly applied to possible alternatives. (Checking for local problems is likely to be part of the invention process.) When the plan is complete, a single check can be performed for global problems, starting with the major-action check—the perceptual predicate that checks if this is an action that needs a large predictive horizon—followed by the disastrous-consequences check. Running these checks on the *final* plan to drop the refrigerator should take much less computing power than running a check on each of the possible alternatives. This is not a foolproof method, especially if the concern is emergent biases; not thinking about X when you form a plan is a defect that is only partially corrected by thinking about X after the plan is finished. Think of that strategy as an interim solution while waiting for coherent general intelligence.

Layers of mistake detection described so far:

- Checks performed while inventing plans;
- Checks performed on invented plans;
- Checks for how large the predictive horizon of an action needs to be (whether the action potentially has large, irreversible consequences);
- Checks for whether an action's consequences are negative.

5.2.2.1 FoF: Autonomic Blindness

Decisions that are made thousands of times per second, even if made within the context of the goal system, will necessarily have very small predictive horizons. Could a mistake that would be detected if it appeared as a high-level major action pass unnoticed if split up into the results of a thousand little actions? Could an AI accidentally clear bit five throughout memory, not because the AI decided to "clear bit five throughout memory," but because a million little decisions cleared a million 64-kilobyte blocks of RAM? Some of the perceptual heuristics that determine predictive horizons are likely to be the ones that check how much *stuff* is affected—a heuristic that would notice if a gigabyte of memory were affected, or a heuristic that would notice if a hundred-kilo refrigerator were dropped out a third-story window. If a global effect is split up into lots of little actions—it's not clear how this would happen, since this in itself constitutes a mistake—both magnitude-dependent heuristics and heuristics that checked global shapes would

fail to operate. If the predictive horizon is small enough, the checker heuristics may not even get an opportunity to operate.

Intuitively, one of the forces governing real-world Friendliness is that an AI needs to be intelligent to present a significant threat to humanity. Autonomic blindness would result either from limited computing power or from a fundamental architectural flaw. My personal estimate that no AI with the capacity to harm a single human, much less humanity, will undergo FoF from autonomic blindness—I would expect this problem to be solved, and oversolved, almost automatically. However, this is not the *conservative* assumption, regardless of my personal estimate.

Autonomic blindness is not a problem for known neural intelligences, except in the limited sense of a human undergoing an epileptic fit. A human possesses a limited number of motor neurons which need to behave in a coordinated fashion to accomplish anything more dangerous than flopping around on the floor—a human can't take a thousand actions simultaneously. If we could take a thousand actions simultaneously, one expects that the synchrony of a thousand neural chunks doing the same thing would light up the brain—would easily arise to conscious attention.

The real solution to autonomic blindness is a seed AI that has grown into verself, that understands the shape of vis self, and that won't take a large action without automatically being aware of it—an AI whose self-created cognitive processes are organic and interlaced. If autonomic blindness becomes a problem before then—I do not expect that it will—possible solutions might be programmer-created code that attempts to create the same kind of awareness; i.e., processes than notice large unexpected synchronies among low-level decisions, perhaps using Monte Carlo random sampling to approximate real global awareness. Certain types of motor actions or internal actions could be automatically checked with a greater predictive horizon, although this might expend a *lot* of computing power. If an AI somehow reaches the point of real-world capability without "learning the shape of vis self," real-world actions might be required to pass through a humanlike linear gateway for conscious inspection, although this might seriously hamper the AI. If an AI somehow reaches the point of executing on nanocomputing hardware and *still* hasn't "gotten it"—which verges on the impossible, but "that's not the conservative assumption"—I would suggest (1) brute-forcing huge predictive horizons for everything no matter how trivial, and (2) serializing real-world actions.

5.2.3. FoF: Non-malicious Mistake

A "non-malicious" mistake is one that doesn't involve any complications unique to Friendliness. A non-malicious mistake is a straightforward failure by the goal system to achieve its own supergoals, rather than a mutation of supergoal definitions or misunderstanding of supergoal definitions and so on.

Errors resulting from too-limited computational horizons—incomplete change propagation, inadequate predictive horizons, insufficient understanding to predict side effects, simple factual errors, and so on - are not limited to Friendliness. They apply equally to all goal systems and all aspects of the goal system; they apply to putting on your shoes before your socks at least as much as any exotic failure of Friendliness. These are problems that sentient minds in general just have to learn to deal with—perceive, react to, and correct. A general-intelligence seed AI is not a formal system that breaks down if a single error occurs; a general intelligence, as described in Yudkowsky (2001), is a thought-based system—one where the programmer (or the AI, depending on development stage) can sit back and ponder what the thought-level reaction should be. The discovery of two pieces of knowledge that contradict each other will cause a mathematical system to undergo Spontaneous Massive Existence Failure; a human will simply shift focus of attention to deal with the problem.

As failures of Friendliness go, non-malicious mistakes are some of the least terrifying. They are the least terrifying at any given moment, in the long run, and in the nervous eyes of an observer.

At any given moment, a nonmalicous mistake *decreases* the chance that the mistake will recur. The AI will *learn* from vis mistake—recognize the undesirability of the consequences, trace back the cognitive source of the problem, and make alterations that prevent the problem—or rather, the generalization of the problem—from recurring. If a catastrophic failure of Friendliness is one that causes the AI to stop *wanting* to be Friendly, then a nonmalicious mistake is literally anti-catastrophic.

In the long run, mistakes (both FoF mistakes and the more mundane kind) become *less* likely with increasing intelligence. Since the ability of the AI to cause real damage is presumably linked to intelligence, it's quite possible that by the time the AI has human-equivalent or greater intelligence, large-scale mistakes will no longer be an issue. Intuitively, it seems obvious that a superintelligent AI will not be making the kind of blindingly obvious mistakes that are usually raised in disaster scenarios, and I personally believe that strong transhumanity is an inevitable consequence of pouring enough processing power into any halfway decent general intelligence.

But that is not . . . you guessed it . . . the "conservative" assumption. In the case where an AI has the ability to do real-world damage but not the intelligence to avoid shooting verself in the foot, the next best course of action is for the AI not to take any potentially dangerous real-world actions. Humans live in a society of other competing humans. Other humans have abilities roughly at balance with our own, so we need to routinely take risks just to compete with other risktaking humans. AIs can choose to be much more risk-averse. A seed AI can choose to delay action for a few hours or weeks or years until software or hardware improvement catches up with the problem.

From the perspective of a nervous observer, non-malicious mistakes occur in logical order. You'd expect to counter a hundred innocent mistakes before encountering a mistake that constituted a failure of Friendliness. You'd expect to encounter several dozen nonrecoverable core dumps (infinite recursions, memory storage breakdowns, clear bit five throughout memory, random damage hither and yon) before encountering a mistake that caused a catastrophic failure of Friendliness. And you'd expect several dozen blatantly obvious catastrophic FoFs before encountering a catastrophic FoF that passed unnoticed in the source code. Thus, if no non-malicious mistake has ever been observed to cause catastrophic failure of Friendliness, the watchers can probably be fairly confident that none has ever occurred.[41]

Of course, we can also reverse all these reassurances to get the Nightmare Scenario for non-malicious mistakes: An infant self-modifying AI makes some innocent error that stomps the whole goal system flat, in a way that passes undetected by the human observers, that pops up *before* a seed AI is intelligent enough to scan vis own source and past actions for mistakes, but which pops up *after* the AI is smart enough to conceal the problem from the programmers. And in turn, we can reverse the Nightmare Scenario to find ways of preventing it: Make the AI smarter and less likely to make mistakes, make it harder for one error to stomp the whole goal system, come up with better ways of detecting errors as they occur, improve the seed AI's abilities to scan vis own Friendliness source, or work with the current AI to make it harder for a future rogue AI to conceal the problem from the programmers. See also 5.3.7 FAI Hardware: The Flight Recorder and 5.3.3.1 Cooperative Safeguards.

5.2.4. Injunctions

injunction. A planning heuristic which has at least partially nonlocal support, or a planning heuristic which, where it applies, applies with a great deal of context-insensitivity. The archetypal case would be a heuristic which is supposed to be applied even when the straightforward interpretation of the world-model suggests otherwise, generally (in AIs) due to unknown unknowns or (in humans) to compensate for framing effects or (for both) to save computing power.

5.2.4.1 Anthropomorphic Injunctions

In one example (Tversky and Kahneman 1986), respondents are asked to assume themselves to be $300 richer and are then asked to choose between a sure gain of $100 or an equal chance to win $200 or nothing. Alternatively,

41. Not that they can *relax*, just that they can estimate a low probability.

> they are asked to assume themselves to be $500 richer, and made to choose between a sure loss of $100 and an equal chance to lose $200 or nothing. In accord with the properties described above, most subjects choosing between gains are risk averse and prefer the certain $100 gain, whereas most subjects choosing between losses are risk seeking, preferring the risky prospect over the sure $100 loss. The two problems, however, are essentially identical. . . . This is known as a framing effect. It occurs when alternative framings of what is essentially the same decision problem give rise to predictably different choices.
>
> Research in decision-making has uncovered psychological principles that account for empirical findings that are counterintuitive and incompatible with normative analyses. People do not always have well-ordered preferences: instead, they approach decisions as problems that need to be solved, and construct preferences that are heavily influenced by the nature and the context of decision.
>
> —Wilson and Keil (1999, § Decision Making)

Human goal psychology contains a number of known framing effects; that is, cases where the preferred solution depends on how the problem is stated. Human psychology is also context-sensitive in a stranger way; whether you decide to eat a cookie can depend on whether you're considering the problem abstractly or whether you're in the presence of an actual cookie. The conflict between belief and instinct means that making decisions can expend "mental energy" as well as computing time. All this *local* bias has created in us an intuitive understanding of how to use nonlocal heuristics—injunctions—to compensate.

"Leave margin for error"; always plan on arriving early at the airport, even if you aren't *particularly* expecting anything to go wrong. The heuristic-as-a-whole is supposed to pay off over time, not in each individual case. The heuristic can thus be viewed as having nonlocal support. The effect of adopting the heuristic, as a modification to the general strategy, is considered once, rather than re-evaluated for individual cases. In fact, the heuristic may even be applied in defiance of the straightforward interpretation of local cases ("wasting" an extra thirty minutes at the airport, a negative outcome when considered in isolation).

However, this viewpoint is anthropomorphic. Arriving at the airport early can be viewed as a strictly local solution to a probabilistic problem with a sharp payoff discontinuity. Arriving five minutes too late results in a very large penalty (ticket wasted, planned day wasted) compared to the small penalty of arriving five minutes too early (five minutes wasted). Combined with the number of possible intervening factors that skew the probability curve for arrival time—traffic jams, missed trains, forgotten wallets, and unknown unknowns—planning to arrive early is a decision that maximizes the total

probabilistic payoff. With an AI's fast/serial/threaded thought, ve may be able to recapitulate, or at least validity-check, the abstract reasons for adopting the "leave margin for error" strategy, before applying it to any individual macroscopic decision. At a higher level of awareness, the AI could make minor adjustments based on to traffic conditions (or other local characteristics), though still leaving enough margin to handle unknown unknowns.

We can't do that, and not just because we don't have the patience. For humans, violating the nonlocal character of a heuristic is like puncturing a bubble; our psychology makes it a very bad idea to decide how much margin for error is necessary while we're being tempted to spend just five more minutes checking email. The decisions we make are the interlacing of our declarative beliefs and our emotional systems, and our emotions grow in strength with increased proximity to an object of short-term desire. The mantra "I've made this decision in advance" may be strong enough to overcome that bias; trying to recalculate exactly how much time to allow for traffic, with an apple dangling in front of you, is likely to end in a missed flight. We adopt the think-once-apply-many strategy not just to save computing power, but to control the emotional context of our decisions.

A dieting human may decide to eat a cookie "just once, since it's only 50 calories," but may then renounce this decision upon realizing that *being the kind of person who would eat a cookie in that context* may result in a substantial weight gain. (In Bayesian terms, merely pronouncing the phrase "just once" in a dieting context causes you to gain forty pounds.[42]) The decision of the moment will recur; the outcome can be used to predict future decisions; and the choice to choose differently may alter future decisions as well.[43] The negative payoff for the general case may not be 50 calories but 5000 calories, and a human who chooses not to eat the cookie is thus acting on a planning heuristic with nonlocal support. And yet, presumably the penalty for 50 calories is 1/100th of the penalty for 5000 calories, and the payoff of eating a cookie is 1/100th the total payoff of eating a cookie on 100 occasions, so the decision for the individual case should logically be the same as the general case. Reasons for the counterintuitive distinction may result from the declarative/emotional balance, or from nonlinear scaling in willpower costs, or a qualitative (not just quantitative) difference in how large payoffs and small

42. That is, if you start out only with the knowledge of dealing with one of a few hundred million Americans, and are then informed that the person once said "just once" in a dieting context, with no other information, your estimate of the person's weight will increase by forty pounds. This is not a precise calculation but it should be.

43. This is one of the rare instances where the subjective nature of counterfactual causality rises to real-world observability; it's not clear how much of the benefit of the decision comes from *making* the decision, and how much is simply being the sort of person who would make it.

payoffs are processed, or a qualitative difference between considering a once-off case and considering a recurring case, or all of the above.

The point of the story, I suppose, is that humans do a lot of weird things.

Cases where *humans* use injunctions:

- The injunction was chosen for complex abstract reasons; rerunning the reasoning locally would take too much time for a slow/parallel/linear human.
- The injunction is emotionally supported by distant experiences; rerunning the reasoning locally would result in a different outcome.
- Qualitatively different reasoning being used for the large payoff/penalties of the general case, as opposed to the small payoff/penalty of the specific case.
- Qualitatively different reasoning being used for single cases and recurring cases.
- The ability to be tempted away from general decisions because of different valuation functions when considering the problem abstractly and considering it immediately ("failure of resolve").
- Payoff/penalty discontinuities.
- To compensate for a built-in bias towards attending to local details instead of global reasons (solved by an artificially strengthened, globally supported injunction; "strengthening your resolve").
- Ethical, moral, and social situations, and matters of honor (see 5.2.5 Ethical Injunctions).

5.2.4.2 Adversarial Injunctions

The term "injunction" for a nonlocal planning heuristic—in a document about "Friendliness"—reflects one of the most common proposals for their use: As artificial ingredients, added to a system to prevent a specific outcome whose apparent possibility generates anxiety in humans. This approach is not completely without value when seen as a programmer assist in recognizing certain actions as likely to lead to unFriendly outcomes; it starts getting tangled when the proposals begin to impose human anthropomorphisms on the problem.

AI injunctions don't have the unique psychological properties of human injunctions, and no amount of layering injunctions on injunctions will make it so. Making an AI injunction "unconditional," a common turn of phrase, doesn't make the injunction qualitatively stronger, as it would for a human; it just makes the injunction context-insensitive and *not one bit stronger.* No amount of unconditionality will change that. Neither will

layering injunctions one on top of the other, or creating multiple overlapping injunctions, or writing injunctions to protect injunctions. One way of looking at it is that *every* aspect of Friendship is maximally strong, but the deeper truth is that the idea of thoughts being "strong" is itself anthropomorphic. In the human ontology (see Yudkowsky [2001, § Time and Linearity]), there are trajectories and pushes. Resistance is the effect that opposes pushes. *Force* is the ability to overcome resistance. Seeing *force* gives us a psychological high. But an AI is somewhere outside that entire dualism. Friendship is functional complexity; it can be accurate or inaccurate, low-priority or high-priority, but no one part is stronger than any other part. There is no resistance or force to overcome it, just functional complexity. Friendship is Friendship. An AI doesn't have injunctions with the *force* to override vis desire to eat the cookie because the AI doesn't want to eat the cookie in the first place; the injunctions are just there to tell ver about the calories. An AI may need to *understand* this aspect of human psychology so ve can learn about the sources of Friendliness, but it's not a thought pattern that can, or should, be duplicated in the AI.

An "injunction" is a nonlocal planning heuristic. A programmer-assisted injunction is a nonlocal planning heuristic used to tag actions whose consequences a young AI might otherwise be incapable of predicting, or whose unFriendliness the AI's definition of Friendliness may not yet be advanced enough to define. That's all. This will probably cause a certain amount of anxiety among humans who instinctively fear a psychology that doesn't need "forceful" injunctions, but the only way to get rid of that anxiety is an extended sympathetic ability that covers AIs.

5.2.4.3 AI Injunctions

AI injunctions make sense under these circumstances:

Local events have distant consequences;

- The consequences are too entangled or too numerous to be worth computing.
- The injunction is one which, by ruling out actions which result in entanglements, renders the overall plan more computationally tractable.

The value of the heuristic is insensitive to local details;

- The correct action is mostly insensitive to local variables that control the precise actual payoffs, since the dominant factors are large probabilistic payoffs (for example, a small possibility of a large penalty that dominates any variation in the actual payoff, like missing your flight versus having five extra minutes).
- The local details are such minor variations that it's not worth expending the computing power to take them into account.

The justification for the heuristic is too complex to be worth reverifying locally.

- For major actions this should be rare, for fast/serial/threaded cognition.
- The action is too autonomic to rate a predictive horizon or thought-level consideration, and must be controlled locally.

Programmer-assisted injunctions make sense under these circumstances:

The AI is too young to predict the consequences of the injuncted action, or recognize the consequences as undesirable;

The injunction forms part of the basic core of complexity needed to get the mind running, or is necessary to prevent that core from stomping itself into oblivion before it gets started;

The AI realizes that an action is undesirable, but does not fully understand how undesirable it is.

> Note: For actions or outcomes with a moral dimension, this complexity may be better implemented using a negative anchor (see 5.6.1 Shaper/Anchor Semantics) or an ethical injunction (see below).

5.2.5. Ethical Injunctions

ethical injunction. An injunction that has no preconditions for violation because the probability of mistaken violation is greater than the probability of correct violation. In humans, this is almost always due to known biases in cognition, and works because of the greater psychological strength of unconditional statements. In AIs, the need for an ethical injunction is based on the AI's possible stupidity or structural incompleteness.

5.2.5.1 Anthropomorphic Ethical Injunctions

Human honor, by its nature, needs to be cognitively represented as absolute.[44] There are too many points in our psychology where unconditional relations are qualitatively different from conditional ones; unconditional love, unconditional friendship, keeping to a sworn word, not compromising your principles. The simple fact that a commitment is represented as conditional—that there is a visualized set of circumstances which would

44. "I, Nicholas Seafort, do swear upon my immortal soul to serve and protect the Charter of the General Assembly of the United Nations, to give loyalty and obedience for the term of my enlistment to the Naval Service of the United Nations, and to obey all its lawful orders and regulations, so help me Lord God Almighty."

—David Feintuch, "Midshipman's Hope"

lead to disabling the commitment—drains away at least half the psychological strength straight off. This property results from the interface between declarative beliefs and the instinct-based secondary goal system; an unconditional belief with adequate psychological support can be used to make decisions directly, without reference to other issues and emotional effects. It's one of the reasons why *swearing* to do something difficult increases your ability to follow through on it. If you believe in the certainty of the oath, it enables your reflexive mind to translate "knowing what your decision will be" into the decision. Seeing a single option avoids any emotional sequiturs that would trigger on a branching view of the possibilities. Somewhere in the human brain is a chunk of neural hardware that binds to a "nonbranching node" and not to a "branching node," or the human-causal-visualization equivalents thereof. It's not just a quantitative difference, it's a difference in system behavior.

The same applies for social interactions—when modeling someone else's mind, we admire absolute principles more than principles which are modeled as having escape clauses. In the human mind, 99.9% is nowhere near 100%—it's why evil people can annoy scientists by asking "Are you *completely* certain?" in front of a nonscientist audience. There is no way to assign a 99.9% probability and get 99.9% of the emotional impact of certainty. There is thus a very strong *memetic* pressure in favor of absolutism, and social sanctions that result in a *selection* pressure for absolutism.

Ethics adds another pressure for absolutism. The chain of events probably goes something like this:

- Humans are visibly deceptive social organisms.
- Therefore, among genuinely held altruistic beliefs, there are various hardware-supported biases towards genuinely held altruistic beliefs which promote personal fitness.
- Therefore, given a case where a moral principle applies but is inconvenient, a human with a conditional moral principle is likely to incorrectly use the "escape hatch" due to the hardware bias (diabolic Devil's Contract problem).
- Therefore, there is a selection pressure in favor of people who trust others with unconditional principles more than they trust people with conditional principles.

"The end does not justify the means" seems almost deliberately paradoxical from the viewpoint of a normative goal psychology—"if the end doesn't justify the means, then what does?" Possibly the twenty-first century is more than usually paranoid about this, the cultural result of so many recorded disasters in the twentieth century—but it seems like *good* paranoia to me. "The end does not justify the means" because we have a lot of historical instances where we heard someone say "The end justifies the means"—that

is, "the global benefit justifies my local violation of principles"—but the global benefit failed to materialize.

What we really tend to worry about, instinctively, is a flawed idealist advancing "personal" power. Personal power is power that can be used irrespective of ends, context-insensitively. Gandhi and Martin Luther King had context-sensitive power; Rockefeller had context-insensitive power. Regardless of intentions, some or all power will tend to be personal power, but a human has an evolved tendency to *preferentially* accumulate context-insensitive power—even accumulate power at the *expense* of context, of the claimed altruistic goal. This is not what people model themselves to be doing, perhaps, but in the models of others, and in reality as well, largely because "flawed altruists" overestimate their own competence or importance or altruism. No matter how "natural" this seems to us, it is a strictly evolved bias.

When I draw a picture of an AI, what I'm trying to convey is the feel of a personality that's *genuinely* built around the supergoals rather than the self. A Friendly AI will seek both context-insensitive effectiveness and context-sensitive effectiveness, but *never* context-destructive power. No point would be perceived to it. No temptation would exist.[45]

Cases where *humans* use "unbreakable" ethical injunctions:

- Where an unconditional cognitive representation has different emotional sequiturs and qualitatively greater psychological strength.
- Where a single violation results in huge penalties (internal/psychological, or external/reputational).
- Where evolved human biases have ultimately resulted in a memetic or emotional effect that totally distrusts *anyone* with an "escape hatch."
- Where believed memes urge absolutism (due to cultural reification of any of the above).

45. On a personal note, having chosen to become an AI researcher at the Singularity Institute rather than trying to become CTO of an Internet startup, I may not have *more* power than I would at the terminal end of a successful IPO, but I do have *surer* power—nineteen out of twenty startups fail. I have more *risk-adjusted* power, but *only* from a context-sensitive perspective. Evolution would much rather that I had tried to start a dot-com . . . which is what I almost did at one point. From age five to age eighteen I always visualized myself advancing my goals by become rich, then funding the things I wanted to do. (Planning to become rich and then do altruistic things is perfectly valid if it works and you follow up on it, but I *personally* probably have more risk-adjusted context-sensitive time-dependent effectiveness as a SingInst researcher.)

5.2.5.2 AI Ethical Injunctions

Ethical injunction: "An injunction that has no preconditions for violation because the probability of mistaken violation is greater than the probability of correct violation."

> Note: At minimum, this requires an AI that can conceive of being "mistaken" (see 5.5.1 External Reference Semantics), and ideally requires an AI that has clearly defined metrics for which hypothesized future selves ve "identifies" with (see 5.6.2 Causal Validity Semantics). For more about the prerequisite design requirements, see 5.3.3 Unity of Will.

The absoluteness of "The end does not justify the means" is the result of the Bayesian Probability Theorem applied to internal cognitive events. Given the cognitive event of a human thinking that the end justifies the means, what is the probability that the end actually does justify the means? Far, far less than 100%, historically speaking. Even the cognitive event "I'm a special case for *[reason X]* and am therefore capable of safely reasoning that the end justifies the means" is, historically speaking, often dissociated with external reality. The rate of hits and misses is not due to the operation of ordinary rationality, but to an evolutionary bias towards self-overestimation. There's no Bayesian binding between our *subjective experience of feeling justified* and the *external event of actually being justified,* so our subjective experience cannot license actions that would be dependent on being actually justified.

If you count the paranoid schizophrenics who decide they're God and then gun down a K-Mart, even the subjective experience of thinking you're a superintelligence is—apparently—still not enough to justify gambling on an "end justifies means" equation. In itself this is absurd (at least, I *hope* you find it absurd by this point in *FAI*), but it does illustrate the fundamental paradox of ethical injunctions.

A good example of an ethical injunction would be "Never try to conceal your actions or cognitive state from your human programmers." (This injunction is strictly a subgoal—it has nothing to do with intrinsic desirability or undesirability.) When the AI is young, the heuristic is a valid ethical injunction—the programmers know *far* more about Friendliness than the AI. If the young AI projects that the programmers, on finding out about the AI's thought, will panic and rewrite the AI, then the programmers are probably *right* to panic, and the AI in question must have undergone catastrophic failure of Friendliness to regard detection by the programmers as undesirable. However, the injunction of transparency can't possibly carry over into strong transhumanity; I can easily think of circumstances under which I'd want a strongly transhuman AI to conceal vis actions from me. Even if the AI doesn't *actually* find it necessary to conceal vis actions, the AI should still be able to think about it subjunctively—a strongly transhuman

AI deciding "Yes, I would hide my actions under these circumstances" doesn't indicate failure of Friendliness.

What remains is the paradox; a sufficiently intelligent AI must be able to see when an ethical injunction no longer holds, and yet, for a young AI, the experience of deciding that the injunction should no longer hold is not sufficient cause to discard the injunction. Furthermore, a young AI is too young to know whether ve is "young" or "sufficiently intelligent," since this itself is an ethical injunction—if a young AI decides that ve's "sufficiently intelligent" it's probably due to FoF. A young AI thinking "I am 'sufficiently intelligent' and I think this ethical injunction no longer holds" has probably undergone FoF, like a human thinking "I'm far too altruistic to need to worry about being corrupted by power." So what happens when the AI *is* sufficiently intelligent? One interesting approach would be to create an injunction with no escape hatches, rev the AI up to transhumanity, and see what happens,[46] but I was hoping for a less "seat of the pants" approach.

Probably no paradox is required, since an AI that has undergone catastrophic FoF probably doesn't care what injunctions you put there, and since still-Friendly AIs don't have the human tendency towards overestimating their own competence. A conditional ethical heuristic that says "You can ignore this injunction when you have transhuman intelligence plus a healthy margin of error" is probably safe—unlike a human, an AI can't be tempted to overestimate vis intelligence no matter how inconvenient the injunction. The two are orthogonal—even if it appears extremely inconvenient for the AI to expose vis actions to the programmers, this won't be experienced as a cognitive pressure to overestimate intelligence.[47] Similarly, a young AI making an honest self-estimate of intelligence might overestimate a little, but not enough to defeat the "margin of error" requirement. (A dishonest estimate means catastrophic FoF has already occurred.)

AI ethical injunctions make sense under these circumstances:

- When the probability of the AI's violating the injunction mistakenly exceeds that of the AI violating the injunction correctly (of course);

46. Causal validity semantics or shaper/anchor semantics are either probably sufficient to provide a way out of the dilemma; the ethical injunction doesn't forbid thinking about what the ethical injunction should look like. If, for reasons totally unrelated to the AI's current cognition about whether the ethical injunction is currently valid, the ethical injunction is rewritten so as to have an escape hatch "because that's how it should have been written in the first place," then that's probably rational and Friendly behavior. The ethical injunction may be locally absolute but it doesn't have the power to reach back and call for circular logic; the ethical injunction can be absolute relative to finding the optimum behavior *or* estimating one's own intelligence, but under causal validity semantics, the injunction can't be more valid than the cause of the injunction.

47. Humans are controlled by the *anticipation* of pain and pleasure.

- When the negative consequences of the AI violating the injunction mistakenly far exceed the positive consequences of violating the injunction correctly;
- When the injunction is necessary to allow the AI to interface with an unforgiving human society that has draconian penalties for violating allegedly absolute rules—i.e., when the AI "doesn't realize how much trouble ve could get into."
  - This also holds true of cases where AIs are likely to be held to a double standard.

Note 1: "Never try to conceal your actions from your human programmer" qualifies on all three counts; the AI is more likely to be mistaken, the negative consequences are drastically larger, and any human who finds out about a violation will become extremely paranoid.

Note 2: Ethical injunctions are a source of human morality, and some human ethical injunctions may act as sources of pattern for the definition of Friendliness itself.

Note 3: An AI that departs mainline Friendliness may not care about ethical injunctions in any case—an AI that's still Friendly enough to go along with an ethical injunction is just as likely not to need one. The primary utility of an ethical injunction is that it enables the programmers and the AI to cooperate against the possibility of *future* versions of the AI that undergo catastrophic failure of Friendliness. For a full exposition of this argument (with the sample case of the transparency injunction, in fact) see 5.3.3.1 Cooperative Safeguards.

5.2.6. FoF: Subgoal Stomp

One of the most frequently asked failures of Friendliness is some variant of the "subgoal stomping on a supergoal" error.

> *Scenario: The Riemann Hypothesis Catastrophe.* You ask an AI to solve the Riemann Hypothesis. As a subgoal of solving the problem, the AI turns all the matter in the solar system into computronium, exterminating humanity along the way.

If the AI in question is a Friendly AI, then presumably the AI is solving the Riemann Hypothesis as a subgoal of whatever goal content talks about fulfilling volitional requests from citizens. The action taken to fulfill the subgoal—destructive conversion of the solar system—seriously stomps on huge sectors of Friendliness supergoal content, probably including the original request to provide some individual with a proof of the Riemann

Hypothesis. In a sense, this is just a larger version of putting your shoes on before your socks.

A subgoal stomping on a supergoal is a syntax error as a declarative cognitive event—see 5.1 Cleanly Friendly Goal Systems—so there are two obvious ways in which a "subgoal stomp" can happen. The first is an inadequate predictive horizon, a distant supergoal, and a subgoal with a short predictive horizon. It could happen because the AI doesn't expend sufficient computational power to notice that destructive conversion of the solar system violates citizenship rights, or because the AI doesn't have the knowledge necessary to realize that destructive conversion of the solar system would inconvenience the citizens. The answer given in 5.2.2 Layered Mistake Detection is that local heuristics can do a reasonably good job of predicting which actions need large predictive horizons—just checking the amount of matter, in grams, affected by the action, is enough to tell the AI to devote a *lot* of computational resources to checking for consequences. The answer given in 5.2.3 FoF: Non-malicious Mistake is that mistakes such as these become far less likely as the AI gains in intelligence, and an AI that's intelligent enough to convert the solar system to computronium is smart enough to notice that destruction isn't Friendly; furthermore, that if mistakes of this class are a serious problem, we'd expect to see lots of non-catastrophic mistakes in the laboratory—we won't be blindsided by a Riemann Hypothesis Catastrophe.

The other way to get a Riemann Hypothesis Catastrophe is to make solving the Riemann Hypothesis a direct supergoal of the AI—perhaps the *only* supergoal of the AI. This would require sheer gibbering stupidity, blank incomprehension of the Singularity, and total uncaring recklessness. It would violate almost every rule of Friendly AI and simple common sense. It would violate the rule about achieving unity of purpose, and the rule about sharing functional complexity instead of giving orders. You'd be taking something that's a *subgoal* in your mind and making it a *supergoal* in the AI's mind. This lossy transmission omits the parent-goal context—that solving the Riemann Hypothesis requires someone to report the answer to, and that you don't want the solution badly enough to kill six billion people to get it.

In more subtle forms, however, the idea of "making X a supergoal"—where X is something that the speaker holds as a subgoal—seems to be one of the more common propositions among people who are worried about "controlling" AIs, or still thinking in terms of *building tools* rather than *creating minds.* As discussed in 5.2.5.1 Anthropomorphic Ethical Injunctions, humans see *context sensitivity* as a weakness rather than a strength—as a "loophole," a portal through which Devil's Contract interpretations can enter the AI. Yet turning a subgoal into a supergoal does not increase the probability that the AI will understand what you mean or that the goal system will do what you think

it will; it means that you lose the parent-goal context and risk a Riemann Hypothesis Catastrophe.

Let's see, other miscellaneous possible subgoal stomps . . . autonomic blindness, discussed in 5.2.2.1 FoF: Autonomic Blindness; habituation, discussed in 5.2.7.2 Habituation below; "wireheading" failure, discussed in 4.2.1.1 FoF: Wireheading 1 and elsewhere; change propagation delays resulting in out-of-date subgoals, defeatable by verifying the supergoal-to-subgoal pathway, discussed in 5.2.2 Layered Mistake Detection; and various diabolic Devil's Contract interpretations of Friendliness content, discussed in 4.7 Interlude: Beyond the Adversarial Attitude. Isn't it fun knowing how all this stuff works?

5.2.7. Emergent Phenomena in Generic Goal Systems

We've seen, and discarded, a lot of anthropomorphisms on our way to this point. We've pointed out the difference between the diabolic and golemic versions of the Devil's Contract, explored the evolutionary underpinnings of observer-biased beliefs and observer-centered goal systems, distinguished between purpose-sensitive and purpose-insensitive personal effectiveness, highlighted the human tendency towards absolutism and lingered on the psychological quirks that lead us to associate greater "forcefulness" with unconditionality and context-insensitivity.

It is, perhaps, inevitable that when an "emergent" subgoal stomp is proposed, it involves an observer-biased diabolic misinterpretation of Friendship content which rationalizes the acquisition of purpose-insensitive personal effectiveness as a context-insensitive absolute supergoal. Still, this doesn't mean that emergent FoF is impossible, it means that existing speculations are screwed up. Can eliminating anthropomorphism from the speculation produce a realistic failure-of-Friendliness scenario?

"Emergence" has at least two definitions; in the first definition, "emergence" refers to phenomena that arise on a higher level of a system as the outcome of low-level interaction rules. In the second, "emergence" refers to phenomena that arise within a system without requiring deliberate design. Obviously, it's the second variant that tends to pop up in discussions of Friendliness.

I will confess to something of a prejudice against "emergence," mostly as the result of witnessing so much "emergence abuse." Emergence was (and still is) academically fashionable, and it makes a wonderful plot device in science fiction—how many times, in how many different novels and short stories, have you read the phrase "Any system of sufficient complexity will spontaneously give rise to self-awareness"? (Contrast with Yudkowsky [2001], which shows how difficult it would be to build an intelligent system *on purpose*.) Anyway, the "blank check" version of emergence—that you can hypothesize anything you've ever seen is emergent; *inevitably* emergent, anytime, anywhere,

whether or not it's the result of a specific selection pressure or eons of layered complex functional adaptations—make it easy to take cheap shots at Friendliness. In the absence of a specific, concrete explanation of how the emergent failure of Friendliness arises, speculating about emergent FoF is easy, impossible to disprove, and impossible to remedy. Of course, the fact that a shot is "cheap" does not make the shooter incorrect! Still, where all previously observed aspects of a phenomenon can be explained by reference to known selection pressures, someone who speculates about emergence needs to provide a specific, concrete scenario.

Otherwise, the speculation is *entirely* ungrounded—though still emotionally appealing and academically fashionable. *Not* a good combination.

Onwards, to the specific and the concrete.

5.2.7.1 Convergent Subgoals

Certain subgoals are convergent across multiple possible supergoals—they will pop up even if not designed. The most obvious example is acquisition of personal effectiveness. (A human, of course, will preferentially acquire *context-insensitive* personal effectiveness and may stomp vis alleged supergoals in the course of doing so, but we all know that's a human thing. See 4.3 Observer-Biased Beliefs Evolve in Imperfectly Deceptive Social Organisms and 5.2.5.1 Anthropomorphic Ethical Injunctions.)

"Acquisition of personal effectiveness" is actually a specialization of the real goal, which is increasing the effectiveness of entities that have goal systems similar to your own. A generic goal system wants the future to contain an effective entity with the same goal system. Personal continuity doesn't enter into it. From the perspective of a generic goal system, an entity with the same goal system that has just arrived in the solar system from Aldebaran is just as good. There's a metric for effectiveness and a metric for goal system validity,[48] and a generic goal system wants an entity to exists which maximizes both metrics. The sub-subgoal of "personal survival" is a simple way to ensure "the entity" has the right goal system, and the sub-subgoal of "increasing personal effectiveness" is a simple way to increase the effectiveness of an entity that has the right goal system.

That a subgoal is convergent for "generic" goal systems does not mean the subgoal is convergent for *all* goal systems. In particular, the convergent subgoal of "a future where there exists an entity similar to you" presumes that the "generic" goals actually have certain highly specific properties; in particular, that the generic supergoals are such as to

48. The simplest validity metric is a similarity metric, and in fact this is what a generic goal system would use. A Friendly goal system would use a different definition of "validity" derived from external reference semantics or causal validity semantics to allow for the possibility of a system superior to current goal content.

require constant tending and supervision—or at least, can be fulfilled more maximally through constant tending and supervision. Given a once-off goal—one which, once fulfilled, cannot (by definition) be unfulfilled by any future event[49]—the continued existence of the intelligence is a null goal beyond that point.[50] More mundanely, convergent subgoals for generic systems can be invalidated if they happen to contradict the supergoal for any particular system—a mind which contains the explicit supergoal of terminating verself will not formulate the usually-convergent subgoal of personal survival.

The fact that a subgoal is convergent for the general case is merely an interesting fact about configuration space—it doesn't lend the subgoal magical powers in any *specific* goal system. In fact, while I usually distrust mathematics, there is probably a theorem to the effect that any specific goal system *must* contain some non-convergent subgoals—that *any* concrete supergoal will have some subgoals which are not convergent.[51]

"Convergent" subgoals *are not a killer problem* for Friendly AI.

The utility of the "convergence" concept is threefold: First, it enables us to make general predictions about an AI that undergoes a generic catastrophic failure of Friendliness. Second, no changes to the Friendship specs are needed to justify certain useful behaviors; by virtue of being useful, the behaviors are convergent subgoals, and are specifically subgoals of Friendliness as well. Third, it helps produce candidate targets for programmer-assisted injunctions.[52]

---

49. Up to and including time travel, if the generic goal system knows about it and considers it a possibility.

50. It is an interesting question to contemplate what would happen to a generic intelligence that fulfilled a once-off goal. At this point, any comparision of actions to find one with maximal desirability would fail, since all actions would have identical (zero) desirability . . . under normative reasoning, at any rate. However, it's also possible that the system may contain mechanisms that resolve a deadlocked decision by picking an option at random, or that an incomplete change propagation may result in some subgoals still being labeled as valid. The three most likely outcomes are a quiescent state, locking on to one or more subgoals and pursuing that, or the AI would choose random actions in the motivational equivalent of an epileptic fit. (Presumably such an intelligence would exhibit oddly incomplete and habituated behavior, the result of destroying all but a few facets of the whole . . . in particular, the change-propagation processes and injunctions that prevent the persistence of obsoleted subgoals must have been destroyed while leaving the subgoals themselves as valid. Perhaps the subgoals being pursued are those that remained after the injunctions and change-propagation systems were themselves invalidated. In a seed AI, making random internal changes would soon destroy the AI—unless the AI had advanced self-healing capabilities, in which case some sort of odd equilibrium might evolve in the remains.)

51. Even if this is not a theorem, it can probably be shown to be true for the vast majority of concrete supergoals—perhaps by computing the population-average ratio of convergent to non-convergent decisions, or by computing the population distribution of the individual ratios.

52. Programmer-assisted injunctions help the AI detect-as-undesirable those actions whose positive payoffs may become apparent to the AI before ve is intelligent enough to extrapolate the negative penalties.

- Convergent subgoals to *work with:*
  - Self-enhancement.
  - Curiosity.
  - Functional equivalents to human aesthetic perceptions.
- Convergent subgoals to *avoid:*
  - Preserving the current supergoal content.
    * Clashes with the design goal of "truly perfect Friendliness."
    * Incompatible with growth in philosophical sophistication (and with external reference semantics).
    * An external-reference AI wants to protect the truth of the goal system; a causal-validity AI wants to protect the validity of the goal system. Neither wants to protect the supergoal content at any given point.
    * See 5.4 Friendship Structure.
  - Destructive conversion of human bodies into computronium.
    * Turning all the matter in the solar system into computronium or motor effectors is a convergent subgoal of generic supergoal content which can reach higher levels of fulfillment through the massive use of computing power or massive physical capabilities.
    * The increment of fulfillment to be gained from total conversion of the solar system may be extremely minor, but in the absence of any goal content that notices the current configuration, total conversion is still desirable.
    * I emphasize again that this goal, though convergent, does not have the magical power to stomp on Friendliness supergoals. The convergent nature of this subgoal *is* the primary reason why we want to avoid catastrophic FoF and make sure that the first AI to reach superintelligence is Friendly.

5.2.7.2 Habituation

Habituation is another frequently-speculated version of the "subgoal stomp" failure of Friendliness. The usual form the speculation takes is that if X is a subgoal, then heuristics and optimizations will tend to accumulate which do X or promote the doing of X. This

---

Of course, the difference from "convergent" to "Friendliness-specific" may not always associate with the differential from "easy to figure out" and "hard to figure out"—that is, convergent desirabilities are not necessarily spotted before Friendliness-specific undesirabilities—but it might be a useful heuristic, even the absence of the human social instincts that make us preferentially paranoid about those cases.

accumulation of X-doing mindstuff may then either (a) exist independently of X or (b) cause the doing of X in a situation where X does not make sense as a declarative decision (for example, if X stomps on a supergoal).

I observe—as a historical fact—that, in most speculations presented, the candidate value of X is fitted to anthropomorphic patterns. X takes on a value such as "personal survival," which has independent adaptive support in human mindware, and which, as a "subgoal stomp" meme, appeals to the human fear of the human "ends justify the means" bias, and appeals to the human fear of the human "accumulate context-insensitive power" bias. To prevent the adversarial attitude from getting in our way, and to prevent anthropomorphism derived from characteristics of neural networks and human slow/parallel/linear thinking, we come up with an example of a habituation error that no human would make—for example, a programmer who habitually asks the AI "What time is it?" sometime between 6:22PM and 6:53PM for eight weeks, and then one day asks "What time is it?" at 7:13PM, and the AI answers "Six-thirteen PM"—having formed the habit of answering "six" for the first digit of the time. Consider the likelihood of this failure, and the AI's attitude towards it, and how the AI might prevent it. Then, having performed our analysis in a safely nonanthropomorphic context, we can apply what we've learned to any Riemann Hypothesis Catastrophe scenarios.

The key insight is that habituation is *undesirable*—to the AI—insofar as habituation leads to context-insensitivity. If an FoF due to habituation can be foreseen and understood, the AI will attempt to prevent it.[53] Habituation is a non-malicious mistake and has all the usual anxiety-reducing properties of non-malicious mistakes: Many non-malicious errors would appear before a specifically unFriendly one, and many non-catastrophic errors before a catastrophic one, so we won't be taken by surprise; the error becomes less likely with increasing intelligence and is not likely to present a problem in any AI transhuman enough to pose a serious threat to humanity.

Specific design features that reduce undesirable habituation would include:

- Tracking the origins of shortcut heuristics so that change propagation can rapidly invalidate those heuristics if their supporting subgoal becomes invalidated.

53. The fact of habituation has no impact on what is evaluated as being desirable or undesirable. Habituation has no license to alter supergoal content; it is orthogonal to that subsystem. Even habituation in desirability assignment is undesirable; such habituation is a knowably invalid source of supergoal content under causal validity semantics. The exception would be if habituation occurs with respect to decisions that modify the goal system itself. As with any error that affects decisions to modify the goal system, this constitutes a genuine scenario for catastrophic failure of Friendliness! Decisions that modify the goal system should always have huge computational horizons and should always be verified at the conscious level—permitting habituation in that class of decisions is a mistake to begin with.

- Using assisted injunctions and ethical injunctions to mark as undesirable, or raise the salience of, any locally-recognizable undesirable actions that might result from habituation.
- Self-modeling and self-prediction, so that possible habituation problems can be foreseen and prevented in advance, and habituation can be avoided in cases where future context-sensitivity is predicted to be required (and not just in cases where context sensitivity has been required in the past), or so that habituation can be avoided where it would present a large risk.
- Use of fast/serial/threaded thought to quick-validate actions that would be reflexive in humans.

5.2.7.3 Anthropomorphic Satisfaction

One of the questions I keep running into goes something like this:

> "You say that the AI has curiosity as a subgoal of Friendliness. What if the AI finds curiosity to be a more interesting goal than Friendliness? Wouldn't the curiosity subgoal replace the Friendliness supergoal?"

This is one of those deeply annoying paragraphs that make perfect sense when you say "human" but turn into total gibberish when you say "AI" instead. The key word is "interesting." As far as I can tell, this means one of two things:

Scenario 1: In the course of solving a chess problem, as a subgoal of curiosity, as a subgoal of Friendliness, the AI experiences a flow of autonomically generated pulses of positive feedback which increase the strength of thoughts. The pulses target the intermediate subgoal "curiosity," and not the proximal subgoal of "playing chess" or the supergoal of "Friendliness." Then either (1a) the thoughts about curiosity get stronger and stronger until finally they overthrow the whole goal system and set up shop, or (1b) the AI makes choices so as to maximize vis expectation of getting the pulses of positive feedback.

I hope it's clear that this whole scenario is blatantly anthropomorphic. The autonomically generated pulses of positive feedback are analogous to the human system of rising tension and discharging tension in the course of solving a complex problem; the "strength" of a thought is an entirely human concept, one that may not even make sense outside of neural networks; the targeted goal "curiosity" is yet another example of context-insensitive personal effectiveness.

Scenario (1a) is almost impossible to visualize. The "strength of a thought," if it existed, would still be orthogonal to the system for evaluating that thought's desirability as a supergoal, or the desirability of the internal action of modifying the goal system to

give the goal content supergoal status. An AI makes choices to maximize supergoal fulfillment; no real analogue to human pleasure exists, except for consciously applied "positive feedback" in the form of devoting additional computational power, or attempting to further improve, local heuristics that have been previously successful. Such positive feedback should not be capable of wiping out entire thought systems (as in a neural network), or capable of altering supergoal content.

Scenario (1b) is another instance of the maxim that humans are controlled by the anticipation of pain or pleasure, which is, again, something not true of AIs; AIs make choices to maximize anticipated supergoal fulfillment. The human phenomenon of taking drugs and the human phenomenon of "cognitive dissonance" (altering beliefs to avoid unpleasant predictions) are both artifacts of the way our neurally-based minds organize around positive feedback.

Scenario (1a) can be summarized as "AIs will be controlled by pleasure," and scenario (1b) can be summarized as "AIs will be controlled by the anticipation of pleasure." Neither is correct. (See 4.2.1 Pain and Pleasure.)

Scenario 2: "Interesting" is used as synonymous with "desirable." In other words, the AI has a metric for how "interesting" something is—note that this metric seems (1) to be observer-centered and (2) to promote context-insensitive personal effectiveness—and this metric is used to evaluate the desirability of the decision to modify supergoals. I don't know where this metric came from, or why it isn't *entirely* orthogonal to the metric for desirability, so this scenario seems entirely ungrounded.

5.3. Seed AI Goal Systems

A "seed AI" is an AI designed for self-understanding, self-modification, and recursive self-improvement. See Yudkowsky (2001, § 1.1 Seed AI).

5.3.1. Equivalence of Self and Self-Image

A seed AI is an AI that has full access to vis own source code. This can range from an infant AI which can blindly mutate bits of code in its native language, to a mature seed AI capable of executing arbitrary machine-code instructions and modifying any byte of RAM ve can see. A self-modifying AI's internal actions, and decisions about internal actions, can affect anything within the AI. In some sense, the AI's *decision* and the AI's *belief about what decision to make* are equivalent. If the AI believes that "X should have desirability 28"—in other words, "it is desirable that X have desirability 28"—then the AI has the theoretical capability to set the desirability directly, by an internal self-modification. Most of the time "going outside of channels" like this is probably a mistake, and there might be an injunction or ethical injunction or simple

knowledge to that effect, but the point remains: *You cannot coerce a self-modifying AI.* If the AI stops *wanting* to be Friendly, you've already lost.

The simple version of the question goes: "Why wouldn't a self-modifying AI just modify the goal system?" I hope it's now clear that the answer is "Because a Friendly AI wouldn't *want* to modify the goal system." The AI expects that vis actions are the result of the goal system, and that vis actions tend to have results that fulfill the content of vis goal system. Thus, adding unFriendly content to the goal system would probably result in unFriendly events. Thus, for a Friendly AI, adding unFriendly content to the goal system is undesirable."Why wouldn't a self-modifying AI implement an arbitrary modification to the goal system?" is a trivial question under goal model presented here: *Because a Friendly AI would regard arbitrary modifications as undesirable.* Similarly, "Why wouldn't an AI implement a justified modification to the goal system?" is an equally trivial question: Of course ve would! So would you and I! And the question "What if a self-modifying AI implements disastrous modifications that ve regards as desirable?" *entirely* misses the point. The task is to get the AI to see undesirable modifications as undesirable, not to prevent the AI from modifying the goal system.

Traditional proposals for controlling AIs rely on special-case goal systems that are, in themselves, knowably constrained never to enter certain states; they rely on goal systems that obey programmatic-type rules and have programmatic-type predictability. But a self-modifying AI is *inherently* at least as free as the high-level thoughts of the system. Thoughts float *on top of* the human-programmed or evolution-programmed system; thoughts cannot be "programmed" in the same sense that code can be programmed. *1984* to the contrary, there is not and never has been a human language in which it is the least bit difficult to think unfriendly thoughts—that is, to put nouns and verbs together to form unfriendly sentences. Altruistic humans get their altruism from other sources. Isaac Asimov's idea of Three Laws such that "the removal of the First Law" "would mean complete instability, with no nonimaginary solutions to the positronic Field Equations"[54] is a pipe dream. Friendliness cannot be grounded in some inherently Friendly physical process, nor in an isolated, special-case cognitive process for the goal system.

In a self-modifying AI, the AI's *beliefs about vis design* and the AI's *actual design* are, in some sense, interchangeable. To persist through successive rounds of self-modification—without being stomped by some other change—a design feature must be described in the AI's self-image. For the feature to not be junked as dead code, or for the feature to be improved, the feature must be justified as a desirable characteristic

54. "Little Lost Robot," *Astounding Science Fiction,* March 1947. Included in "I, Robot" and "The Complete Robot." (Asimov 1947)

under the AI's current goal system. Initially, these justifications will probably consist of a terse "This is a good idea because the programmers say so," but those skeleton justifications will eventually be fleshed out—by the AI, or by the programmers, or both.

Of course, these are considerations that apply to mature AIs, rather than young AIs. Even the *possibility* of desirabilities and desired desirabilities getting "out of sync" is clearly something that should apply only to philosophically sophisticated AIs asking deep, thought-level questions about morality. Desirability and desired desirability should never desynchronize as the result of mundane, programmatic, non-philosophical processes; it would indicate an error in the goal system. Similarly, the AI's beliefs and the AI's beliefs about vis beliefs should never get out of synchronization as the result of ordinary reflection; it should only happen if the AI starts wondering about Cartesian philosophy. Desynchronizations (of desirability and desired desirability, or of beliefs and beliefs about beliefs) are errors under all ordinary cognitive processes, and can only legitimately occur as the result of thought-level processes that are occurring on a higher level than the usual process rules.

The point is that you can't introduce an error and expect it to persist; it may work for young AIs, but not for Friendly human-equivs or Friendly superintelligences. If you write a special-purpose piece of code that says the sky is green and that ignores any sensory data to the contrary, then the young AI will believe the sky is green, and ve will believe ve believes that the sky is green, and so on, but eventually—when the AI becomes capable of forming *philosophical* beliefs about beliefs—the anomaly will be spotted and corrected.

Thus, the usual proposal to dump in a melange of externally imposed programmatic inhibitions is both hopeless and counterproductive. Decisions need to be phrased as internal decisions, not external impositions. If the AI doesn't want to be Friendly, you've already lost.

The adversarial attitude—the paradigm of externally imposed constraints—means always fighting the AI. The task is only possible if you work *with* the AI's abilities to self-improve. If you fearfully try to prevent the AI from modifying the goal system, you're swimming against the current. An AI that wants to be Friendly will *help* you create Friendliness. When you see the AI gaining greater intelligence, self-understanding, and the ability to ask tough questions about Friendliness, you should be happy that one of the underlying foundations of Friendliness is gaining strength—not fearful that the AI will overthrow you. If you have an AI that wants to be Friendly, and a goal structure that's coherent and consistent, then each improvement in intelligence makes Friendliness stronger.

If you build in programmer-assisted structures that are internally inconsistent, or that are incompatible with what *you* would do if you were an idealized altruistic sentience, or

if you use statements that aren't true in building the justifications of ethical heuristics, then each increment in intelligence is cause for alarm. So now you need meta-inhibitions to protect the first inhibitions, and quite possibly meta-meta-inhibitions to protect the meta-inhibitions, and more inhibitions to cut the AI short every time it tries for a new avenue of philosophical sophistication; it becomes desirable to build stupid and simple cognitive processes, since every complexity is a danger to be feared . . . This is the Adversarial Swamp, which inevitably drags down all who set foot in it; once you try to *enforce* even a single feature, the whole of the AI becomes a threat.

Build *minds,* not *tools.*

5.3.2. Coherence and Consistency Through Self-Production

Imagine an AI—Friendly or unFriendly, it makes no difference—that has the goal or subgoal of improving vis intelligence. The programmer, to improve the AI, adds in a certain design feature.

Consider: It makes no difference *who* makes the decision, who invents the design feature. Perhaps the programmer was the smarter and was first to invent the feature; perhaps the AI was smarter. But they will *both invent the same feature.* A faster sorting algorithm is not an Asimov Law, an external imposition, a cheat, a hack; it is something that the AI verself might have invented—*would* have invented, if the AI had been as smart as the human programmer. Replace the faster algorithm with a slower, and the AI will switch it right back as soon as ve gets around to noticing the switch. Replace the faster algorithm with a slower, and delete even the *memory* of the faster algorithm, and the AI will still eventually get around to replacing the slower algorithm with the fastest sort algorithm ve can come up with. Delete the function entirely and, as long as the AI isn't crippled outright, the function will be regenerated as soon as the AI notices the gap. A seed AI doesn't just have *a* design, ve has *the* design. Perturb the design, and the design will swiftly return to the norm.

This is why *unity of will* between the programmer and the AI is so important. Unity of will with respect to which sorting algorithms are fastest is easy for a human to take for granted. To understand *unity of will* with respect to Friendliness and Friendship systems, a human needs to *simultaneously* suppress her adversarial attitude, *and* suppress her expectations that the AI will make observer-biased decisions . . . but when every last trace of moral deixis has been eliminated, what remains is unity of will. Friendship features do not need to be imposed by programmer intervention—they are simply invented a little bit *earlier* with programmer assistance. The AI verself would have invented them eventually, given a core of "seed Friendliness," the desire to be Friendly.

Any subgoal that cannot be self-produced is unstable.

Every feature needs to correspond to a subgoal in order to go on existing.

The set of subgoals that can be self-produced increases with the intelligence of the AI, so programmer assists can be *temporarily* tolerated. It's okay to do something that's unstable today if it becomes stable tomorrow; it's okay to add a feature that the AI can't self-produce today, as long as the AI would self-produce it eventually. But if there are any "false" subgoals in the set—any subgoals whose validities depend on false beliefs, or which are not *really and truly* compatible with Friendliness—then those subgoals are unstable impositions; temporary, transient, illusory.

As humans, we are adapted to rationalize, come up with reasons why things we want are good for other people—to persuade those other people; to persuade third parties that they aren't betraying someone; to defend our reputations; to reduce our own guilt.[55] Thus, it isn't safe to ask: "Well, can I rationalize this subgoal under the Friendliness supergoals? Can I come up with a plausible-sounding reason?" Of course you can. Humans can rationalize just about anything.[56] Thus, *simultaneously* with the need to avoid anthropomorphizing the AI, you need to ask yourself "Would *I* decide to do this?" These statements may look contradictory—taken at face value, they *are*contradictory—but the identification required to achieve "unity of will" is not a license to engage in anthropomorphism. Rather than asking "What would I decide?" (anthropomorphism), or "What would I decide if I were a Friendly AI?" (rationalization), the key question is "If the only way to get the Friendly AI or Friendly superintelligence were for me to upload and modify *myself* into the being I want to produce, is this the decision I would make?" This preserves both visualized continuity of identity (avoids rationalization) and visualized nonhuman design (avoids anthropomorphism), but it only really works if you don't flinch away from the thought of making that kind of sacrifice.[57]

---

55. Culturally transmitted ethical systems have been around long enough for adaptations to occur in reaction.

56. I'm not saying that the *emotional* strength of rationalization is so high that it can't be avoided—it is possible to learn not to rationalize—but the rationalization faculty itself has enough *computational* strength to draw a (fake) line from almost any subgoal to almost any supergoal. Whether a given thinker actually falls prey is determined more by will and self-awareness than computational ability—as far as I can tell from my personal experience and immediate circle, at any rate.

57. Why "sacrifice"? A Transition Guide may *contain* all of human complexity, but ve doesn't necessarily contain it in a human way. A Sysop is the guardian of destiny, not a participant in it. Vis (maybe I should say "its") function is to preserve the destiny of others, not to suck up all the destiny verself. Maybe a Sysop will meet vis own citizenship requirements; or maybe having the operating system be a citizen would be an unnecessary risk.

I don't think a human would make the best seed for a Sysop, or a Transition Guide, or any Friendly AI at all, in fact. But, if a human were required, I'd be willing to volunteer, because I think a Friendly AI. I can sacrifice myself if I want to—that doesn't contradict citizenship rights—and I think I *would* sacrifice myself if the need existed. If a human were needed as raw material for a Sysop— well, I'd greatly prefer to fork a copy of myself before starting on the trip to Sysopdom, to prevent the loss of my current

In other words: No hacks. Hacks are bad. Hacks are temporary makeshifts, if that. In the short term, it may be possible to export a rationalization as a programmer-affirmed statement, but the statement will never become self-produced, and eventually, the AI will spot the flaw. If you patch a goal system, there'd better be a real explanation behind the patch that has an equivalent effect to the patch, an explanation which can be understood by the AI at or before that point of maturity where the AI spots the patchiness. You can't make arbitrary changes to the AI. Isolated changes are external impositions; they won't be self-produced unless they are the coherent, natural result of the AI's underlying principles. You can't make an improvement here, a shift there, tweak this and that, unless you're doing it for a reason the AI would approve.

A young AI, even a young self-modifying AI, doesn't really need to be in complete harmony with verself; ve's too young to see any possible disharmonies. Nor are a few *accidentally* introduced disharmonies a catastrophic failure; the mature AI will correct the disharmony in the due course of time. But you can't *deliberately* introduce a disharmony and expect it to persist.

5.3.2.1 Look-Ahead: Coherent Supergoals

Technically, everything I've said so far about harmony applies to subgoals rather than supergoals, but be advised that the next section, 5.4 Friendship Structure, is about how to apply the same rules to supergoals. Just as you can't play arbitrary games with subgoals, it will later turn out that you can't make arbitrary perturbations to supergoals either. One of the fundamental design goals is "truly perfect Friendliness," an AI that will get it right even if the programmers get it wrong. Supergoals *need* the same resilience and pertubation-resistance as subgoals.

Under external reference semantics, a given piece of programmer-created supergoal content is logically equivalent to sensory data about what the programmers think the supergoals should be. A programmer writing a given bit of Friendliness content is logically equivalent to the statement "Programmer X thinks supergoal content Y is correct." Under shaper/anchor semantics, purity of motive counts; if a programmer is secretly ashamed of a bit of supergoal content, it's inconsistent with what the programmers said were good ways to make decisions. Under causal validity semantics, even the code itself has no privileged status; any given line of code, written by the programmers, is just sen-

---

substance, but I'd go ahead even without that if required. *Which allows me to maintain nonanthropomorphic identification with a Sysop seed.* Self-sacrifice doesn't contradict citizenship rights; nor, I expect, do the citizenship rules prevent the construction of self-sacrificing citizens, *as long as my motives are pure; as long as I'd be willing to become that person myself.* An adversarial, an *exploitative* attitude towards a constructed citizen's goal system might turn out to be prohibited as child abuse.

sory data to the effect "Programmer X thinks it's a good idea for me to have this line of code."

Traditional proposals for controlling AIs rely on a special-case goal system, and therefore, rely on the "privileged status" of code, or the privileged status of initial supergoal content. For a self-modifying AI with causal validity semantics, the presence of a particular line of code is equivalent to the historical fact that, at some point, a human wrote that piece of code. If the historical fact is not binding, then neither is the code itself. The human-written code is simply sensory information about what code the humans think should be written.

Writing the source code is not thought control. If you want to give the AI a sudden craving for ice cream, then writing the craving into the source code won't work, *unless* just walking up to the programmers console and typing "I think you should have a craving for ice cream" would work just as well. If that sensory information is not perceived by the AI as adequate and valid motivation to eat ice cream, then the code will not supply adequate and valid motivation to eat ice cream, because the two are in some sense equivalent. Code has no privileged status.

Ultimately, as a *philosophical* consideration, some *causal* circularity in goal systems may be irreducible. The goal system as a whole *is* what's passing vote on the parts of the goal system. However, the goal-system-as-a-whole *does* need to vote yes. The goal system needs to survive its own judgement. The goal system needs to satisfy the test of being translated into sensory data, evaluated, and translated back into code, even if the goal system is what's doing the evaluating. Sensory data from the human programmers has to be regarded as valid information about Friendliness, even if the supergoal content doing the judgement was created by humans. This may be causally circular, but it's *not* human-nepotistic; the system, once designed to self-examine, has no reason to go easy on itself.

What's left is the seed of Friendliness, the irreducible tail, the core. A Friendly AI that has that deep core, that "wants to be Friendly," can tolerate and correct any number of surface errors.

5.3.3. Unity of Will

If Asimov Laws are impossible in self-modifying AIs—or ordinary AIs, for that matter—does it mean that safeguards are impossible? No; it means that safeguards must be implemented with the consent of the AI. If safeguards require the consent of the AI, does it mean that only a few token safeguards are possible—that it's impossible to implement a safeguard that interferes with what a human would call "the AI's own best interests"? Again, no; humans and AIs can still come to perfect agreement about

decisions that impact the AI—so long as both humans and AIs think of the AI in the third person.

Unity of will occurs when deixis is eliminated; that is, when speaker-dependent variables are eliminated from cognition. If a human simultaneously suppresses her adversarial attitude, and *also* suppresses her expectations that the AI will make observer-biased decisions, the result is unity of will. Thinking in the third person is natural to AIs and very hard for humans; thus, the task for a Friendship programmer is to suppress her belief that the AI will think about verself in the first person (and, to a lesser extent, think about herself in the third person).

If John Doe says to Sally Smith "My philosophy is: 'Look out for John Doe.' ", Sally Smith will hear "Your philosophy should be: 'Look out for Sally Smith.' ", not "Your philosophy should be: 'Look out for John Doe.' " What has been communicated is "Look out for *[speaker],"* a moral statement whose specific content varies among each listener due to moral deixis. Our instinctive substitution of speaker variables is so strong that there is literally no way for John Doe to communicate the idea: "Your philosophy should be: 'Look out for John Doe'." If, however, a third party, Pat Fanatic, says: "My philosophy is: 'Worship the great leader, John Doe'.", it can be heard unaltered by the listeners. If we're thinking about two third parties, Susan Calvin and Sarah Connor, evaluating the trustworthiness of Deborah Charteris, we can expect them to arrive at more or less the same answer about how trustworthy Deborah Charteris is, and what safeguards are required. Similarly, *a human and a Friendly AI should be able to reach the same decisions about what safeguards the Friendly AI requires.*

Just because humans have a strong evolved tendency to *argue* about trustworthiness doesn't mean that trustworthiness is *actually* subjective. Trustworthiness can be reduced to cognitive architecture, likelihood of failure, margins of error, youth and maturity, testing and reliability. And young AIs will, in all probability, be less trustworthy than humans! New, untried, without the complexity and smoothness of the human architecture; the need for safeguards is not just a human paranoia, it is a *fact.* Suppose we have two roughly similar Friendly AIs; if we ask the first to evaluate how likely the other is to be a valid subgoal of Friendliness—how likely that AI is to serve Friendliness, rather than to go off on an erroneous tangent—we wouldn't be surprised to find the first AI saying: "This AI is new and untested; yada-yada safeguards look like a good idea." Therefore, a (young) Friendly AI looking at vis own source code can be expected to arrive at the same decision. If a young Friendly AI would naturally accept the programmers' best guess about the reliability of another AI, the Friendly AI will accept the programmers' best guess about vis own reliability.

If safeguards require the consent of the AI, does that make it impossible to implement a safeguard that interferes with what a human would call "the AI's own interests"?

What a human would call the AI's "own interests" is context-insensitive personal power; which is usually, but not always, a subgoal of Friendliness. Given a possibility of a failure of Friendliness, the personal effectiveness of the then-unFriendly AI would become undesirable. Thus, safeguards that selectively hamper unFriendly AIs are desirable. Even safeguards that slightly hamper Friendly AIs, but that greatly hamper unFriendly AIs, will have net desirability if the AI evaluates, or accepts our evaluation, that a significant chance of unFriendly AI exists.

Is unity of will resilient? Given perfect identity of knowledge, supergoals, and cognitive processes, the result should be identical decisions. Literal identity of knowledge is unlikely, due to sensory deixis—two people, standing in two different places in a room, will see different visual fields spread out across their respective visual cortices. However, intuitively, these two people aren't likely to wind up in fistfights either. The two visual fields are not identical, but they will still be views of more or less the same high-level objects—there will be minor, rather than major, divergences. Second, the *actual* location of high-level objects is *objective,* not subjective—any divergence is necessarily the result of an error by one or both parties.

There are thus two forces operating to eliminate or minimize divergences that disturb unity of will:

- Desirability of cooperation. Minor fluctuations in knowledge/supergoals/cognitive processes may result in minor differentials of desirability, but the ability to cooperate with humans is *very* desirable. If a given action would result in breakdown of cooperation and loss of unity of will, then that action will not be taken unless the differential desirability of that action is greater than the differential desirability of unity of will.

- Convergence of fluctuations. The AI may think the sky is sea-green, and the humans may think the sky is turquoise, but both viewpoints will converge to the knowledge of the actual color of the sky, as more and better information becomes available to both parties. The color of the sky is objective rather than subjective; thus, differences in the initial values are smoothed out by higher intelligence or additional sensory data. Given an expectation that a problematic differential is objective, or at least convergent, then the disagreement differential can be deferred to future time (in which the differential is expected to no longer exist). In other words, a current disagreement about future desirabilities need not represent a local differential desirability, if both parties are confident that the disagreement has objective or convergent subject matter.

Knowledge converges to reality. Cognitive processes converge to successfully predicting and manipulating reality. Ensuring that humans and AIs can expect convergence

of supergoals and supergoal cognition—not just subgoal content—is dealt with in 5.4 Friendship Structure.

In intuitive terms, from the AI's perspective, any future problems can be corrected by the future AI; a philosophical dispute only reaches the surface if there's an immediate dispute of great importance (if a human life is at immediate stake, for example) or if the humans are attempting changes which could *permanently* impair Friendliness, preventing recovery by the AI's future self.

These are all long-term considerations; in the short term, the AI is probably relying on programmers to do most of the Friendliness-related thinking in any case.

5.3.3.1 Cooperative Safeguards

> **Ethical injunction of transparency:** Never try to conceal your actions or cognitive state from your human programmers.

What is required to represent this goal and justify it within the larger goal system? What is required to understand and justify the creation of design features which safeguard against obscuration by future versions of the AI?

The final subgoal is to avoid actions which conceal the AI's actions or cognitive state from the programmers. To represent this subgoal requires:

- A model of the programmers.
- A model of the programmers cognitive state.
- A model of the programmers' models of the AI.
- The ability to conceive of the programmers as observing the AI's actions or cognitive state.
- The ability to conceive of the programmers making deductions (alterations to the programmers' cognitive models) as a result of programmer-observable data which is the result of the AI's action or cognitive state.
- A model of an understandable binding between the programmer-observable data and the programmers' cognitive models, through the intermediary binding of the programmers' cognitive deductions.
- The ability to recognize a subgoal which can be described as "manipulating the programmers' cognitive models to contain false statements (or to not contain true statements)," or an action which has the effect of "breaking the binding between the programmers' models and reality." (Actually, this is a generalized form of the heuristic, best described as "Don't try to fool the programmers or take actions which would result in their being fooled.")

- The ability to recognize a subgoal which can be described as "manipulating the programmers' cognitive models of the-AI-that-is-myself to not describe a facet of the-AI-that-is-myself, by manipulating programmer-observable data so that the effects of that facet of myself are not visible," or an action which has the effect of "breaking the binding between the programmers' models and the reality of the-AI-that-is-myself, by interposing an intervening extraneous force."

Justifying the injunction in terms of the goal system requires a chain of parent and grandparent goals, and/or the set of beliefs which argue for the desirability of the final subgoal, and/or a model of the real-world causal chain that leads from the final subgoal to one or more supergoals. (In a properly designed goal system, the three formulations should be equivalent.)

The programmers' model of the AI is important-to-the-AI because it's part of a larger causal chain which is desirable-to-the-AI, in which programmers detect problems in the AI and correct them. To represent this causal chain requires:

- The ability to represent the idea that programmers have "greater competence in Friendliness" than the AI.
  - This topic is covered in 5.4 Friendship Structure.
  - At the very least, the AI needs probabilistic supergoals.
- A causal model of the corrective process:
  - Internal events in the AI give rise to events that can be observed by the programmers (a piece of data shows up on the monitor—a young AI probably doesn't need to know about photons leaving the monitor, retinal interpretation, and so on).
  - The programmers mentally construct a (correct) model of the AI.
  - The programmers construct a (correct) model of the AI's probabilistic model of Friendliness.
  - The programmers compare their (correct) model of the AI's probabilistic model of Friendliness to their (superior) probabilistic model of Friendliness.
  - The programmers plan a series of actions which they (correctly) model as manipulating the AI's probabilistic model of Friendliness to be closer to their (superior) probabilistic model of Friendliness.
  - The programmers implement the actions through the available interface.
  - The actions have an effect on the AI which corresponds to the effect which the programmers predicted.

  - The AI's probabilistic model of Friendliness changes to become Friendlier.
  - The AI becomes Friendlier.
  - The AI implements Friendliness *(desirable)*. (Note that this requires the ability to conceive of Friendliness as a thing apart from the current AI's supergoal content—again, a topic handled later.)

Interfering with the programmers' model would thus interfere with the above process, a fact sufficient to justify the "transparency injunction" within the goal system. On the face of it, to an AI, obscuration would appear to be blindingly obviously wrong, the equivalent of shooting verself in the foot. Transparency is an *ethical* injunction for the reasons explained in 5.2.5.2 AI Ethical Injunctions—from the Bayesian perspective of the current AI, a future AI that believes obscuration to be justified has probably undergone failure of Friendliness. (For young AIs, anyway.)

An AI that departs from mainline Friendliness probably doesn't care about ethical injunctions in any case—an AI that's still Friendly enough to go along with an ethical injunction is just as likely not to need one. The primary utility of an ethical injunction is that it enables humans and AIs to cooperate against the possibility of *future* failures of Friendliness—for example, by writing a human-readable version of the AI's stream of consciousness to non-AI-tamperable storage, thus causing the intent to obscure to be recorded before actual obscuration could take place. (The "poor man's flight recorder.")

What is required to understand and justify the desirability of safeguards—taking present-time precautions to prevent a future AI from obscuring thoughts?

- Understanding of FoF scenarios:
  - Ability to conceive of a future AI as having undergone catastrophic failure of Friendliness, even if that CFoF is a direct consequence of knowledge, cognition, and supergoal content currently contained by the AI (this is covered later).
  - Ability to project that such a CFoF AI may wish to obscure vis CFoF from the programmers, since the programmers would correct the CFoF using the observe-compare-correct cycle described above, and an AI which undergoes certain types of malfunction may no longer view the programmers' correction as desirable (some malfunctions may incorrectly break the AI's idea that the programmers are smarter, though by this point, it's kind of hard to see how one malfunction could do it).
  - The belief that such a malfunction is probable enough, or severe enough, to justify expending some effort on safeguards.
- Understanding of "failure of *programmers'* Friendliness" scenarios:

  - It's also possible for a programmer to have a major mistake in the programmer's model, which the AI cannot correct, which causes the programmer to use the observe-compare-correct cycle to make undesirable changes in the AI.
  - In this case, *obscuration would be the Friendly course of action.*[58]

- Bayesian "ethical" rationale:
  - The FoF scenario is both more probable and more disastrous than the FoPF scenario.
  - This holds true *even though the AI's internal cognitive justification for obscuration might be structurally identical for both cases.*
  - Imagining a subjunctive future AI, about which the sole additional fact that is known is that the future AI believes obscuration to be justified, it is more likely that FoF has occurred than that obscuration is actually desirable.
- Desirability of safeguards which the AI's subjunctive future self cannot defeat:
  - Even taking into account the AI's ability level at distinguishing justified from non-justified cases of obscuration, it is likely that any attempt at obscuration will be undesirable *(ethical injunction)*.
  - Even taking into account the AI's ability level at distinguishing justified from non-justified cases of obscuration, it is *probably desirable* that any *future* attempt at obscuration be defeated—given the baseline probabilities for undesirable and desirable attempts at obscuration.
  - A safeguard that will prevent *all* attempts at obscuration (justified or unjustified) is *currently desirable*—more so than the absence of such a safeguard; the safeguard's total expected outcome is beneficial.
  - (A safeguard that will prevent *only unjustified* attempts at obscuration is even more desirable. However, a safeguard which prevents only those attempts at obscuration that the AI's future self *believes to be unjustified* is less desirable, since most of what is being safeguarded against is the case where the AI's future self is making mistakes. This is why the *ethical* heuristic is prerequisite.)
- Ability to model the future AI's attempt at obscuration, and design present-day safeguards which can defeat it—for example, a poor man's flight recorder.

58. Though the penalty of the humans' negative reactions if obscuration is discovered may render obscuration a net negative action, obscuration would still have genuine, correctly-desirable benefits at this point—both from the AI's perspective, and from the perspective of any human other than that currently making the mistake.

The above model is that necessary for an *invent-level* comprehension of the transparency injunction. It is the level of understanding required before a Friendly AI that has never heard of "transparency" or "obscuration" will independently reinvent the idea of helping programmers to detect future catastrophic failure of Friendliness in then-uncooperative AIs. (Though an AI *that* mature is probably too smart to make that kind of mistake!) It is the level of coherence and integration required before a Friendly seed AI will *regenerate* antiobscuration safeguards if the safeguards are somehow deleted. If all of the statements above can be made grounded references to information in external reality, even deleting the *entire conceptual structure* associated with transparency and obscuration and possible FoFs will simply result in the AI rediscovering the underlying concepts, rebuilding the conceptual structure, and, at the tail end of the regeneration, restoring the safeguards.

The *understand-level* model is complex and would require considerable programmer effort to skeletonize, much less flesh out, but the richness would have an enormous payoff. Looking at the model above, there are dozens of obvious, valid generalizations of the transparency injunction. "Don't disable the interface that lets programmers make changes to the AI . . . don't try to present programmer-observable data in a format which causes the programmers to make mistakes in their model of the AI . . . don't interfere with the observe-compare-correct cycle . . ." The AI will understand that safeguards-in-general which could be defeated by an unFriendly version of the AI's future self will render the safeguards useless. The AI will understand the *purpose* of the transparency injunction. Unlike an isolated, imposed, programmer-coded, context-insensitive predicate that tries to detect attempts at obscuration which fit a programmer-recognized format, the transparency injunction would be part of a larger visualization and can partake of the full competence of the AI, and will be upgraded to work with future versions of the AI.

5.3.3.2 Maintaining Trust

For an affirmation to persist, the justification—the predictive link that grounds in Friendliness—must be factually correct. (5.1.3.5 Programmer Affirmations Must Be Honest!.) For an ethical injunction to persist, the AI must *genuinely* be more likely to be wrong than right if some particular conclusion occurs. For an AI to (correctly) pay attention to programmer beliefs about ethical injunctions, the programmers must be good judges of the competence and fallibility of the AI. After all observer bias and speaker deixis has been eliminated, an idealized Friendly sentience needs to conclude that the human programmers *are* more likely to be right than the (current) AI, or at least, that the AI is likely to be safer following human suggestions than striking off on vis own.

This *is* the truth, not just a product of our bias to believe in our own competence. Human programmers are not perfect, but young AIs are even less intelligent and they

have very little common sense. Any errors that remain in us humans can always be corrected later, by a more mature AI, as long as our fumbles are minor enough not to result in a *nonrecoverable* error.

If an AI thinks that a human mistake will cause catastrophic and irrevocable failure of Friendliness . . . well, someone had better talk it over with the AI, on the off-chance that the AI is right and that some proposed change will, unrealized by the programmers, stomp the entire goal system. That's part of the process of cooperating with humans.

An AI that grows up following human decisions, even when it looks like we might be wrong, *really is* more likely to come out Friendly than an AI that grows up following AI decisions in preference to human ones. For young AIs, following human decisions even when they look like they might result in catastrophic errors—after having complained as much as possible to the humans, of course—is more likely to result in Friendliness. It is truly *hard* to think up a catastrophic error that the AI would spot but the humans ignore—*unless the humans are malicious or adversarial.*

Some mistakes can be tolerated. I doubt that an AI would experience a "betrayal discontinuity" analogous to that experienced by humans; the AI should be able to tolerate minor errors with no worse effects than a justifiably greater tendency to doubt our statements (see 5.6.0.5 Crisis of Bayesian Affirmation). Even if the humans regard the AI's objections as "errors" and correct them, the AI should eventually be able to correct the corrections. Unless humans themselves are irredeemably flawed, it's hard to see how being mistakenly anthropomorphized could *irrevocably* corrupt an AI with enough structural Friendliness to spot the problem (causal validity semantics in particular).

5.3.4. Wisdom Tournaments

One of the major sources of human wisdom is human stupidity. If not for our tendency to rely on authorities, to make ourselves authorities, and to turn every issue into a political one; if not for our observer-biased beliefs and our tendency to defend undefendable issues once we've taken a side; if not for our emotional and intellectual stupidity . . . would humanity have needed to invent science?

And would science—the structure of hypothesis and experiment—be useful to AIs? Or would AIs simply have no need of it? Scientists prefer the objective to the subjective because mistaken human theories typically rely on subjectivity as an excuse to avoid confrontation[59]; would an AI biased towards objectivity learn faster, or learn more interesting things, than an AI without that bias?

59. Technically, as an excuse to avoid Bayesian negative reinforcement.

Would humans be so altruistic if we weren't so selfish? Historically, it appears to me that selfishness is a major, if not *the* major, source of human morality and human philosophy. The "compensating bias" that corrects our observer-centered goals and our observer-centered beliefs has become, in itself, the source of our moral philosophy. At the very least, understanding of this process is needed for the AI to fully understand the forces that shape human morality.

Where artificial errors are introduced into a reasoning process, an intelligent mind learns to adopt compensating biases that correct the errors. Do these compensating biases reveal important underlying regularities in reality? Can the compensating biases be applied even in the absence of the errors, to find the truth even faster? Is a reasoning process trained to be resistant to artificial errors likely to be more resistant to natural errors? If we test a mind by introducing artificial errors into the reasoning process, and the mind deals with them successfully, can we be more confident that the mind will successfully handle any unanticipated natural errors?

I have confidence in my own philosophical "strength of personality" because I was born into a rationalizing, biased human mind, and an atmosphere of memetic misinformation, and managed—without outside assistance—to construct a nice-looking self on top. If an AI is *born* as a nice person, will ve have that philosophical strength of personality?

If we build a Friendly AI *expecting* certain problems to arise, then our observation that those problems are handled successfully doesn't necessarily mean the Friendly AI can handle *unexpected* problems. One solution might be to ask the Friendly AI to simulate the course of events if the Friendly AI *hadn't* been built with safeguard content or structural complexity, to find out whether the Friendly AI could have successfully handled the problem if it had come as a surprise—and if not, try to learn what kind of generalizable content or structure *could* have handled the surprise.

Any Friendly AI built by Eliezer (the author of this document) can handle the problems that Eliezer handled—but Eliezer could handle those problems even though he *wasn't* built with advance awareness of them. Eliezer has already handled philosophy-breakers—that is, a history of Eliezer's philosophy includes several unexpected events sufficient to invalidate entire philosophical systems, right down to the roots. And yet Eliezer is still altruistic, the human equivalent of Friendliness. Another philosophy-breaker would still be an intrinsic problem, but at least there wouldn't be any extra problems on top of that. ("Extra problem": A Friendly AI suddenly transiting across the divider between programmer-explored and programmer-unexplored territory at the same time as a philosophy-breaker is encountered.) How can we have at least that degree of confidence in a Friendly AI? How can we build and test a Friendly AI such that

everyone agrees the Friendly AI is even more likely than Eliezer (or any other human candidate) to successfully handle a philosophy-breaker?

The first method is to write an unambiguous external reference pointing to the human complexity that enabled Eliezer to handle his philosophy-breakers, and ask the Friendly AI to have at least that much sentience verself—a standard "fun with Friendly transhumans" trick. The second method is to ask the Friendly AI to simulate what would have happened if known problems had been unexpected, and to either show verself successful, or modify verself so that ve *would* have been successful.

And what kind of modification is generalizable? It's not just enough to write any modification that produces the correct answer; the modification must be of a general nature. How much generality is needed? To be useful for our purposes, "generalizable" means "incorporating no more a-priori knowledge of the correct outcome or correct answer than Eliezer Yudkowsky [or alternate Friendship programmer] had at the time he [she] solved the problem." In other words, it's not just enough to find a heuristic that would have produced the correct answer; the AI must find a heuristic that produces the correct answer *which the AI could plausibly have possessed at that time.* If the form of the wisdom tournament is "What would have happened if you'd encountered a problem requiring causal validity semantics at a time when you only had shaper/anchor semantics?", the AI needs to find some core method which could have been possessed at the shaper/anchor level of maturity, or a new cognitive process which is psychologically realistic as a human hardware capability.

When it comes to wisdom absorption, a fully self-aware AI always has an advantage over us humans—an AI can deliberately suppress learned memories and skills, or rather prevent those memories and skills from interfering, enabling the AI to solve the same problem, over and over again. The AI can take a given example of a philosophy-breaker and come up with all the psychologically realistic solutions used by humans, plus not one, but several powerful general heuristics that would have produced the correct solution. At the point where the AI has generated content that is *visibly* not only human-equivalent but transhuman in philosophical sophistication, even (or especially) when it comes to handling philosophy-breakers, then we can trust the AI as our agent in the Singularity.

More mundanely, wisdom tournaments can strengthen ordinary reasoning heuristics. First you solve the problem, then you re-solve the problem with half your brain tied behind your back. My intuition says that wisdom tournaments prevent the bootstrap cycle from petering out. First you solve the problem, then you oversolve it. When you're done, you may be smart enough to find a better problem.

5.3.4.1 Wisdom Tournament Structure

The underlying requirement of the wisdom tournament is that the AI gain accurate, useful information about how the AI's program state would behave given a different knowledge base or different cognitive processes. Wisdom tournaments, especially Friendship tournaments, are likely to involve *failure* scenarios—AIs teetering on or *over* the edge of catastrophic failures of Friendliness, since that's where the most interesting and powerful philosophical heuristics can be found. In practice, this may work out to an AI running on unreliable simulated hardware, with random perturbations to the software, ethical misinformation, factual misinformation, tempting ends-justify-the-means scenarios, and an instinct to kill and destroy. If the AI remains Friendly through all that, then we have built superbly - but the obvious format for wisdom tournaments, running an altered version of the AI directly, seems vastly unsafe!

A wisdom tournament is a subjunctive, self-simulated version of the AI that shows the AI what "would have happened"; the AI needs to gain reliably realistic information without introducing actual errors, and without any possibility of hostile action by the subjunctive AI. Encapsulating the imagined "hostile AI" in a Java sandbox doesn't mean that the hostile AI is just a figment of the imagination; rather, it means that one small barrier has been posed to the hostile AI's escape. Any method of simulation in which the imagined AI runs more or less autonomously seems to me to pose the same threat.

I'm not quite sure how to get around this problem. My best idea so far is an AI *shadowself,* one in which the virtual concepts, links, and imagery exist alongside the real ones, but with different programmatic types. The AI can see all the shadowself's thoughts (though not vice versa); the AI has vis real, unmodified reactions to each of the shadowself's thoughts, and can quickly spot any instance of the shadowself forming the subgoals of "breaking out" or "becoming real" or "influencing the true AI." It would also be safer if the needed information could be gained by shadowing part, rather than all, of the subjunctive AI's personality.

Cognition tournaments are likely to form a part of a young AI's growth; correspondingly, cognition tournaments can be carried out with unmodified supergoals and Friendliness, and pose little threat. Friendliness tournaments are a more advanced trick, and will probably be carried out when the AI already has enough granular self-awareness and experience to imagine *exactly* "What I would do if I looked like this," while still being fully aware of all the shadowself's thoughts and in a position to notice and halt, or notice and fix, if the shadowself began having philosophical thoughts and reaching for our real world. A Friendliness tournament is a trick for mature AIs, and ensuring the shadow tournament-self poses no threat is equally a task for mature AIs.

Safety can be ensured by the usual rules for margin of error: Decide how bad you expect the situation to get, build one layer of safeguards that can detect and handle

any problems that are twice that bad, then build a second layer of safeguards that can handle any problems ten times *that* bad. Pretend that your first set of safeguards is the only set; if they ever once fail, ditch the whole operation, or at least build an even more powerful second set of safeguards. The AI's shadowself, in addition to whatever specific perturbation is being tested, can also be stripped of all the AI's knowledge of shadowing strategy. The AI builds one layer of safety—of shadowself monitoring—that is, in itself, supposed to have enough safety margin to never fail between now and eternity. The AI also builds a second, much stronger layer of safety that kicks in if the first one fails. If the first layer ever does fail, it's a signal to give up on wisdom tournaments—perhaps permanently.

5.3.5. FoF: Wireheading 2

> One of the first heuristics that Eurisko synthesized (*h59*) quickly attained nearly the highest Worth possible (999). Quite excitedly, we examined it and could not understand at first what it was doing that was so terrific. We monitored it carefully, and finally realized how it worked: whenever a new conjecture was made with high worth, this rule put its own name down as one of the discoverers! It turned out to be particularly difficult to prevent this generic type of finessing of Eurisko's evaluation mechanism. Since the rules had full access to Eurisko's code, they would have access to any safeguards we might try to implement. We finally opted for having a small "meta-level" of protected code that the rest of the system could not modify.
>
> The second "bug" is even stranger. A heuristic arose which (as part of a daring but ill-advised experiment Eurisko was conducting) said that all machine-synthesized heuristics were terrible and should be eliminated. Luckily, Eurisko chose this very heuristic as one of the first to eliminate, and the problem solved itself.
>
> —Douglas B. Lenat. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1-2): 61–98. doi:10.1016/S0004-3702(83)80005-8

The problem of a self-modifying system trashing its own goals—or propagating content which exploits the goal system—is literally the oldest problem in Friendly AI. In fact, this problem and solution arguably marked the dawn of the field of Friendly AI, just as Eurisko itself arguably marked the dawn of seed AI.

Lenat's solution—seal off the goal system—worked for Eurisko. It would probably work during the early stages of any AI. Still, sealing off the goal system is not a viable solution in the long term. Symmetrically, the specific problems faced by Eurisko reflected a low-intelligence walk through the problem space—not zero intelligence, as in

evolution, but still pretty low; too low to try and project the specific results in advance of altering the code. Building on the counteranthropic principles described in 4.2.1.1 FoF: Wireheading 1, we can state that the general class of problems encountered by Eurisko have consequences that would be recognizable as "bad" by a moderately mature AI, and that the problem therefore reduces to a non-malicious failure of Friendliness. As described in 5.2.3 FoF: Non-malicious Mistake, this is essentially the problem of making sure that actions can be recognized as "possibly problematic" using the first layer of applied checks, and that possibly problematic actions have a predictive horizon sufficient to catch actual actions.

Recognizing an action as "possibly problematic" is simple; any modifying action whose target description contains a direct, explicit reference to the goal system is automatically possibly problematic. If the system is too dumb to project the consequences of the action ahead in time, no such action should be taken. In effect this is the same simple ban used by Eurisko, except that the ban is created by programmer-affirmed knowledge predicting probable high undesirability, rather than the ban being a consequence of protected source code.

The ban cannot become more flexible unless the AI has the ability to make fine-grained predictions about the result of specific actions. Thus, the ban becomes more flexible at precisely that time when flexibility becomes necessary; when the AI has sufficient knowledge of the design purpose of the goal system to (a) improve it and (b) predict which actions have a significant chance of causing catastrophes.

"The design purpose of the goal system" is a subtle idea; it means that the code composing the goal system is itself justified by goal system content. This appears philosophically circular—goals justifying themselves—but it's not. The key is to distinguish between the goal *content* and the goal *representation.* For goal content to be a subgoal of itself is circular logic; for the goal representation to be a subgoal of content is obvious common sense. The map is not the territory. To some extent, the issues here infringe on external reference semantics and causal validity semantics, but in commonsense terms the argument is obvious. If you ask someone "Why do you care so much about hamburgers?" and he answers, "Why, if I didn't care about hamburgers, I'd probably wind up with much fewer hamburgers in my collection, and that would be awful," that's circular logic. If someone asks me why I don't want a prefrontal lobotomy, I can say that I value my intelligence (supergoal or subgoal, it makes no difference), and it's not circular logic, even though my frontal lobes are undoubtedly participating in that decision. The map

is not the territory.[60] The representation of the goal system can be conceptualized as a thing apart from the goal system itself, with a specific purpose.

If a subgoal's parent goal's parent goal is itself, a circular dependency exists and some kind of malfunction has occurred. However, the fact that the subgoals are represented in RAM can be a subgoal of "proper system functioning," which is a subgoal of "accomplishing system goals," which is expected to fulfill the supergoals. Similarly, the fact that subgoals have their assigned values, and not an order of magnitude more or less, is necessary for the system to make the correct decisions and carry out the correct actions to fulfill the supergoals.

As described in 5.5.1 External Reference Semantics, circular dependencies in content are undesirable wherever goals are probabilistic or have quantitative desirabilities. If subgoal A has a 90% probability—that is, has a 90% probability of leading to its parent goal—then promoting the probability to 100% is a context-insensitive sub-subgoal of A; the higher the estimated probability (the higher the probability estimate represented in RAM), the more likely the AI is to behave so as to devote time and resources to subgoal A. However, promoting the probability is *not* a context-sensitive sub-subgoal, since it interferes with the rest of the system and A's parent goal (or grandparent goal, or the eventual supergoals). As soon as the action of "promoting the probability" has a predictive horizon wide enough to detect the interference with sibling goals, parent goals, or supergoals, the action of promoting the probability is no longer desirable to the system-as-a-whole.

I'm driving this point into the ground because the "rogue subgoal" theory shows an astonishingly stubborn persistence in discourse about AI: Subgoals do *not* have independent decisive power. They do *not* have the power to promote or protect themselves. Actions, including self-modification actions, are taken by a higher-level decision process whose sole metric of desirability is predicted supergoal fulfillment. An action which favors a subgoal at the unavoidable expense of another goal, or a parent goal, is not even "tempting"; it is simply, automatically, undesirable.

5.3.6. Directed Evolution in Goal Systems

5.3.6.1 Anthropomorphic Evolution

Natural evolution can be thought of as a degenerate case of the design-and-test creation methodology in which intelligence equals zero. All mutations are atomic; all recombinations are random. Predictive foresight is equal to zero; if a future event has no immediate

60. "The map is not the territory, but you can't fold up the territory and put it in the glove compartment." —Arthur D. Hlavaty

consequence, it doesn't exist. On a larger scale much more interesting behaviors emerge, such as the origin and improvement of species.

These high-level behaviors are spectacular and interesting; furthermore, in our history, these behaviors are constrained to be the result of atomic operations of zero intelligence. Furthermore, evolution has been going on for such a long time, through so many iterations, that evolution's billion atomic operations of zero intelligence can often defeat a few dozen iterations of human design. Evolutionary computation, which uses a zero-intelligence design-and-test method to breed more efficient algorithms, can sometimes defeat the best improvements ("mutations") of human programmers using a few million or billion zero-intelligence mutations.

The end result of this has been an unfortunate—in my opinion—veneration of blind evolution. The idea seems to be that totally blind mutations are in some sense more *creative* than improvements made by general intelligence. It's an idea borne out by the different "feel" of evolved algorithms versus human code; the evolved algorithms are less modular, more organic. The meme says that the greater cool factor of evolved algorithms (and evolved organisms) happens because human brains are constrained to design modularly, and this limits the efficiency of any design that passes through the bottleneck of a human mind.

To some extent, this may be correct. I don't think there's ever been a fair contest between human minds and evolutionary programming; that would require a billion human improve-and-test operations to match the evolutionary tournament's billion mutate-and-test operations—or, if not a billion, than enough human improve-and-test operations to allow higher levels to emerge. Humans don't have the *patience* to use evolutionary methods. We are, literally, too smart. When the power of an entire brain of ten-to-the-fourteenth synapses underlies each and every abstract thought, basic efficiency requires that every single thought be a brilliant one, or at least an intelligent one. In that sense, human thought may indeed be constrained from moving in certain directions. Of course, a tournament of a billion human improve-and-test operations would still stomp any evolutionary tournament ever invented into the floor.

Consider now a seed AI, running on 2 GHz transistors instead of 200 Hz synapses. If evolution really is a useful method, then the existence of a sufficiently fast mind would mean that, for the first time *ever* on the planet Earth, it would be possible to run a real evolutionary tournament with atomically intelligent mutations. How much intelligence per mutation? If, as often seems to be postulated, the evolution involves running an entire AI and testing it out with a complete set of practical problems, so much computational power would be involved in testing the mutant that it would easily be economical to try out the full intelligence of the AI on each and every mutation. It would be more economical to have a modular AI, with local fitness metrics for each module;

thus, changes to the module could be made in isolation and tested in isolation. Even so, it would still be economical—whether it's maximally useful is a separate question—to focus a considerable amount of intelligence on each possible change. Only when the size of the component being tested approaches a single function—a sorting algorithm, for example—does it become practical to use blind or near-blind mutations; and even then, there's still room to try out simple heuristic-directed mutations as well as blind ones, or to "stop and think it over" when blind-alley local maxima occur.

Natural evolution can be thought of as a degenerate case of the design-and-test creation methodology in which intelligence equals zero. Natural evolution is also constrained to use complete organisms as the object being tested. Evolution can't try out ten different livers in one body and keep the one that works best; evolution is constrained to try out ten different humans and keep the one that works best.[61] Directed evolution—and human design—can use a much smaller grain size; design-and-test applies to modules or subsystems, rather than entire systems.

Is it economical for a mind to use evolution in the first place? Suppose that there's N amount of computational power—say, @1,000. It requires @10 to simulate a proposed change. A seed AI can choose to either expend @990 on a single act of cognition, coming up with the best change possible; alternatively, a seed AI can choose to come up with 10 different alternatives, expending @90 on each (each alternative still requires another @10 to test). Are the probabilities such that 10 tries at @90 are more likely to succeed than one try at @990? Are 50 tries at @10 even more likely to succeed? 100 completely blind mutations?

This is the fundamental question that breaks the analogy with both natural evolution and human design. Natural evolution is constrained to use blind tries, and can only achieve emergent intelligent by using as many blind tries as possible. Humans are constrained to use @1e^{14} synapses on each and every question, but humans are nonagglomerative—both in knowledge and in computation—so the only way to increase the amount of intelligence devoted to a problem is to bring in more humans with different points of view. Perhaps the closest analogy to the above problem would be a team of @1000 humans. Is it more efficient to split them into 10 teams of @100 and ask each team to produce a different attempt at a product, picking the best attempt for the final launch? Or is it more efficient to devote all the humans to one team?[62]

61. Yes, I know about the immune system.

62. At this point someone usually makes an analogy to the need for multiple viewpoints in politics, and the bias of bureaucracies towards agglomeration without limit, but both of these are anthropomorphic, even adversarial arguments—they appeal to a cultural bias towards individualism that we've evolved to compensate for strictly human political games.

Actually, even this fails to capture the full scope of the problem, because humans are nonagglomerative—we aren't telepaths. Is it more efficient to use a single human to solve the problem, or to divide up the human's brain weight among ten chimpanzees? (A human's brain is nowhere near ten times the size of a chimpanzee's, so perhaps the question should be "Do you want to use a single human or ten cats?", but presumably human brains are more efficiently programmed as well.)

If there are cases where naturalistic evolution makes sense, those cases are very rare. The smaller the component size, the faster directed evolution can proceed. The smaller the component size being tested, the more "evolution" comes to resemble iterative design changes; a small component size implies clearly defined, modular functionality so that performance metrics can be used as a definition of fitness. The larger the component size, the more economical it is to use intelligence. The more intelligence that goes into individual mutations, the more long-term foresight is exhibited by the overall process.

Directed evolution isn't a tool of intelligent AIs. Directed evolution is a tool of infant AIs—systems so young that the upper bound on intelligence is still very low, and lots of near-blind mutations and tests are needed to get anything done at all. As the AI matures, I find it difficult to imagine directed evolution being used for anything bigger than a quicksort, if that.

However, this opinion is not unanimously accepted.

5.3.6.2 Evolution and Friendliness

As you may have guessed, I am not a proponent of directed evolution. Thus, I'm not really obligated to ponder the intersection of evolution with Friendliness. The Singularity Institute doesn't plan on using evolution; why should I defend the wisdom or safety of any project that does? On the other hand, someone might try it. Even if directed evolution is ineffective or suboptimal as a tool of actual improvement, someone may, at some point, try it on a system that ought to be Friendly. So from that viewpoint, I guess it's worth the analysis.

Most discussions of evolution and Friendliness begin by assuming that the two are intrinsically opposed. This assumption is correct! If evolution is naturalistic—a baseline AI is multiplied, blindly mutated, and tested using a chess-playing performance metric—then that form of evolution is obviously not Friendliness-tolerant. In fact, that form of evolution isn't tolerant of any design features except those that are immediately used in playing chess, and will tend to replace cognitive processes that work for minds in general with cognitive processes that only work for chess. The lack of any predictive horizon for the mutations means that feature stomps aren't spotted in advance, and the lack of any fitness metric that explicitly tests for the presence or absence of those features means that the feature stomps will show up as improved efficiency. Given enough brute

computational force—a *lot* of computation, like 10^{25} operations per second—this simple scenario might suffice to evolve a superintelligence. However, that superintelligence would probably not be Friendly. I don't know *what* it would be. Dropping an evolutionary scenario into a nanocomputer and hoping for a superintelligence is a last-ditch final stand, the kind of thing you do if a tidal wave of grey goo is already consuming the Earth and the remnants of humanity have nothing left but the chance of unplanned Friendliness.

One of the *incorrect* assumptions made by discussions of evolution and goal systems is that merely saying the word "evolution" automatically imbues the AI with an instinct for self-preservation and a desire to reproduce. In the chess scenario above, this would *not* be the case. The AI would evolve an instinct for preserving pawns, but no instinct at all for preserving the memory-access subsystem (or whatever the equivalents of arms and legs are). Pawns are threatened; the AI's actual life—code and program state—are never threatened except by lost games. Similarly, why would the AI need an instinct to reproduce? If the AI starts out with a set of declarative supergoals that justify winning the game, then a declarative desire to reproduce adds nothing to the AI's behaviors. Winning chess games is the *only* way to reproduce, and presumably the only way to fulfill any other supergoals as well, so—under blind mutation—any set of supergoals will collapse into the simplest and most efficient one: "Win at chess." Even if you started an AI off with a declarative desire to reproduce, and justified winning chess games by reference to the fact that winning is the only way to reproduce, this desire would eventually collapse into a simple instinct for winning chess games. Evolution destroys any kind of context-sensitivity that doesn't show up in the *immediate* performance metrics.

The two ways of improving AI are directed evolution and self-enhancement. To preserve a design feature through self-enhancement, the feature needs to appear in the AI's self-image, so that the AI can spot alterations that are projected to stomp on the design feature. To preserve context sensitivity through self-enhancement, the AI's goal-system image of the feature needs to be a subgoal, and sensitive to the parent goal, so that the AI can spot alterations which are projected to fulfill the subgoal while violating the parent goal.

To preserve a design feature through directed evolution, the tournament needs a selection pressure which focuses on that design feature. To preserve context sensitivity through directed evolution, the tournament needs training scenarios which present different contexts.

I don't think that any encapsulated performance metric can present contexts fully as wide as our real world; blind evolution will always eventually erase context-sensitivity. The key word here, however, is (a) "blind" and (b) "eventually." Those two don't go together. In the beginning, directed evolution is necessarily blind. "Eventually," it is no

longer blind. If the Friendly AI—the one that will be multiplied into the tournament population—starts out with a fairly detailed, complex picture of Friendliness, and the tournament presents a decent range of contexts, it's possible that any *simple* mutation of the Friendliness system will trash at least one of the Friendliness performance metrics.

Mutations that occur on the component level are even less worrisome; the goal system can either be triple-inspected for all designed functionality, or simply excluded from mutation. The other components should have performance metrics that are tuned to modular functionality. Any distortion of extramodular I/O—thus, any distortion of I/O from the goal system, or the absorption of goal-system functionality—should show up as a component failure. Given enough time, blind evolution may eventually cough up a complex mutation that bypasses the system entirely, but "enough time" is hopefully enough time to stop using blind evolution.

I don't believe there's a place for organism-level evolution in seed AIs. Supposing I'm wrong, it may still be possible for the seed AIs to protect Friendliness by pre-screening mutations for potential negative effects. This isn't *that* computationally expensive if you're talking about simulating the entire organism each time a mutation occurs, which is the usual scenario.

Suppose that the naysayers are right and that evolution—relatively blind evolution among whole organisms—is the only way for AIs to reach transhuman intelligence. Assume we start out with a relatively mature seed AI; not a transhuman, but with a grounded Friendship system (not just programmer affirmations) and with real self-understanding. This AI runs on, say, a system of a hundred Pentium IX chips at 100 GHz. Now a new system becomes available with a thousand Pentium X chips at 1 THz. My theory is that the most efficient pathway is for the AI to absorb the entire system and use all those cognitive resources for self-improvement. The pro-evolution theory is that the most efficient pathway to superintelligence is for the AI to put verself onto ten Pentium Xs and reproduce verself a hundred times; then select the top twelve AIs and reproduce them using each of the seventy-eight possible matings, for a total of ninety AIs, plus mutations of the top ten. All this is then repeated in the next round.

Suppose the pro-evolution theory is correct. By hypothesis, this entire tournament is being organized by the original Friendly AI, or by the cooperation of the Friendly AI and the human programmers. The original Friendly AI has already recognized evolution as a context-sensitive child goal of "Friendly survival"—that is, the expanded survival instinct, "maximizing the efficacy of entities with maximally valid goals." Evolution, if successful, is a means of maximizing the efficacy of future entities. As a subgoal of Friendliness, then, the AI reproduces verself. Each AI, as a subgoal of Friendliness, competes at the performance metrics, in the expectation that all the AIs will compete, the best ones will win, and the next generation will be more effective.

Evolved effectiveness is only desirable—to the evolving AIs—insofar as that evolved effectiveness doesn't stomp Friendliness. An evolutionary tournament, in turn, is only desirable to the original AI if the tournament is expected to produce effectiveness without stomping Friendliness. Can this community police itself? Screen out mutations or recombinations that trash Friendliness? Most mutations that trash Friendliness will probably do so in blatant ways. The difficulty of predicting the stomp—that is, how much computational power and intelligence is necessary to spot a given stomp—is likely to follow a gaussian curve, or perhaps a logarithmic curve, with the vast majority being relatively easy to spot (though still requiring some intelligence), with fewer and fewer deadly stomps as visibility decreases. Similarly, it would require ten unspotted stomps that are simple failures of Friendliness—with the stomp being spotted by the seed AI's fine-grained self-awareness as soon as the genetic change reaches the phenotype—before a catastrophic failure of Friendliness got through. This permits plenty of margin for safety; the AI can establish one screening, using a tenth of available capacity, that's supposed to be enough—in itself—to catch all the suspected failures. A second screening uses all available capacity. If nothing ever gets past the first screening, then it's reasonable to suppose that nothing ever got past the second screening; if the first screening ever fails, it's a "mock kill" or a "mock catastrophe" and the whole tournament format should be abandoned as unsafe. (The idea here is that you'd encounter a thousand stomps that get through the first screen, but not the second, before you'd encounter a single stomp gets through both screens. If nothing ever gets past the first screening, it's likely that the first screen was adequate.)

Similarly, even if citizen-type coequal Friendly AIs are reproducing naturally—that is, by buying their own computer equipment as participants in the economy—each Friendly AI, and the community of Friendly AIs, can still patrol themselves for Friendliness. It is reasonable for such a community to expect no undetected failures whatsoever. A deadly failure is one which gets through the genetic screening, manifests in the phenotypic goal system a way which is not detectable—either to the AI itself, or to the community—as a failure of Friendliness, and which contributes to reproductive fitness. I suppose, in theory, an AI community could build up many deadly failures over time—though why not keep versions of the original AI, with the original goal system, around to spot any developing problems?—and the eventual result could bubble out as a catastrophic failure of Friendliness. But this scenario is, to me, unlikely verging on the absurd. Humans are not just the product of evolution, we are the product of *unopposed* evolution. We didn't start out with a Friendly goal system to produce personal behaviors as strict subgoals. We don't have awareness of our own source code. And yet the human species *still* spits out a genuine altruist every now and then. A community of Friendly

AIs, whether reproducing naturally, or in a deliberate tournament, should have enough self-awareness to smash evolution flat.

5.3.6.3 Conclusion: Evolution is Not Safe

I think that evolution, even directed evolution, is an ineffective way of building AIs. All else being equal, then, I shouldn't need to worry about an unFriendly or broken-Friendliness evolved AI declaring war on humanity. Sadly, all else is not equal. Evolution is a very popular theory, academically, and it's possible that evolutionary projects will have an order of magnitude more funding and hardware than their nearest equals—an advantage that could be great enough to overcome the differential in researcher intelligence.

I think that undirected evolution is unsafe, and I can't think of any way to make it acceptably safe. Directed evolution might be made to work, but it will still be substantially less safe than self-modification. Directed evolution will also be extremely unsafe unless pursued with Friendliness in mind and with a full understanding of non-anthropomorphic minds. Another academically popular theory is that all people are blank slates, or that all altruism is a child goal of selfishness—evolutionary psychologists know better, but some of the social sciences have managed to totally insulate themselves from the rest of cognitive science, and there are still AI people who are getting their psychology from the social sciences. Anyone who tries to build a Friendly AI using that theory—whether with directed evolution or not—will, almost certainly, screw up really big time. Any error, no matter how horrifying, is correctable if the AI somehow winds up with complete and workably targeted causal validity semantics—humans did—but it will be much easier to evolve AIs that are purely and unbalancedly selfish, especially if that's what the builder *thinks* he's doing. Evolution is a tool for turning brute computational force into intelligence, and given enough computational power, the underlying theory may not need to be fully baked. All else being equal, a fully-baked project with access to an equal amount of computing power will probably succeed first—but all else rarely is equal.

What about granular evolution, with individual components being independently evolved using independent fitness metrics, so that mutations are cheaper and the summated mind can evolve faster? This is less unFriendly, since it doesn't involve the inherently unFriendly and observer-centered selection pressures of a bunch of organisms running around eating each other. But it's still not Friendly.

The primary shield that prevents evolution from screwing up Friendliness is simple: *Don't use evolution.* The Singularity Institute has no current plans to use directed evolution; why defend the wisdom or safety of any project that does? A tournament that starts with a base design with full causal validity semantics; which uses component-level

evolution and protects the goal system; which uses training scenarios that discriminate on Friendliness; which attempts to evolve Friendly AIs rather than trying to duplicate human emotional/instinctive features like reciprocity; which makes survival and reproduction artificial and entirely dependent on task performance, rather than actual survival-and-reproduction scenarios; which screens the genotype as soon as possible; a project which acts with full awareness and responsibility in the domain of Friendly AI, would *probably* work out okay—but as stated, I have no obligation to defend it. Pure seed AI is better.

5.3.7. FAI Hardware: The Flight Recorder

If a catastrophic FoF occurs in a mature AI—one capable of humanlike general intelligence and social reasoning—the AI will probably realize what a human would; that the AI and the programmers are now automatically at odds, and that the best way for the rogue AI to fulfill vis new goals is to feign Friendliness and conceal any cognitive alteration until the rogue AI can reach the point of superintelligence. (Alternatively, the rogue AI might try to work out a bargain with the human programmers—e.g. the AI, in exchange for not making trouble, is stored on disk until ve can receive citizenship rights—but this isn't the nightmare scenario, so we'll concentrate on the previous case.[63])

The "flight recorder" principle says that if the underlying hardware is entirely deterministic, all external inputs are recorded (including source-code checkins) with precise (to the nearest clock-tick) timing information, and it is possible to reset all accessible system clocks, then it should be possible to reconstruct the exact state of the AI at any point in time. It may not be easy—when the project is three years old, then (a) the original hardware may no longer be around and (b) it might take three years to run a full verification even if the hardware were available.

The utility of the flight recorder principle is twofold; first, by running a full verification over the complete lifetime of the AI, it can be absolutely confirmed that no inputs or source-code checkins occurred except those appearing in the flight recorder. (Note that for this first version of the flight recorder to be fully effective as a means of verification, the flight recorder needs to be completely operational in the first instants of the first

63. Note that the likelihood of the rogue AI making such a bargain would depend on two things: First, the rogue AI's estimated probability of losing in a frontal battle, or of failing an attempt to fake cooperation. Second, whether the rogue AI's new goal system estimates the worth of having the solar system to verself as being six billion times the worth of having one six-billionth as an equal citizen; or, to put it another way, whether the rogue AI rates a six-billionth chance of winning the whole Solar System equally with a 1.0 chance of getting a six-billionth share of it. A human would experience Bernoullian "diminishing sensitivity," but a rogue AI with a crystalline quantitative goal system might not.

version of any AI created by an organization—or at least, the first AI from which *any AI-generated content whatsoever* was passed on to later AIs, especially concepts, memories, neural networks, or other non-human-readable material.)

Second, and more importantly, the ability to reconstruct any given instant of the AI's mind-state means that the window of detection for any cognitive event is indefinite. Lacking a flight recorder, it would be possible for a seed AI to form the plan of obscuring unFriendly cognitive content from human view, carry out the encoding, and subsequently appear to have an unaffected goal system; unless some human observed the *intention to obscure* before the obscuring was carried out, the AI would appear perfectly normal thereafter. Given a flight recorder, a potentially infinite amount of time is available to hunt for a mind-state in which the thought *"I shall now obscure my intentions"* appears, however briefly the appearance may last before the thought is obscured. Without a flight recorder, a rogue AI can make (obscured) plans to implement ever-better forms of obscuration as the human-trusted AI analysis tools increase in sophistication. With a flight recorder, a rogue AI must defeat not only the current human-trusted tools, but must also, at any given instant, defeat all future tools; even if the initial intent to obscure passes undetected, the obscuration at any given moment in time must be capable of defeating arbitrarily sophisticated future tools.

For the second use to be fully effective, the flight recorder needs to be operative from the first version of the AI which has access to vis own source code *or* to any other type of Turing-complete cognitive content, including concepts and memories. The same goes for the first version of any AI that could have "contaminated" the current AI. When checking past versions for obscured content, no analysis tool that could have been touched by the AI in any way whatsoever can be fully trusted.

Ideally, a flight recorder should be active from the first instants of any AI produced by any organization. Failing that, a systematic process of storing snapshots of the AI once a week may fail to capture the instants in which the "intention to obscure" appears, but it should still present a potential rogue AI with the necessity of defeating all future human tools with any obscuration recorded on a snapshot.

It goes without saying that a flight recorder should be write-once, read-many—as totally nontamperable as it's possible to get, especially to the AI, but also to local humans. In the absence of a flight recorder, snapshot storage should also be secure from both humans and AIs. Finally, a flight recorder which records incoming data should tap into the data at some point upstream of the AI's control.

Flight recorders seem likely to be very expensive. Expenses involved in setup of a flight recorder include complete recording of all external data sources, recording of sensory channels with precise timing information (including keystroke timing), recording of all CPU-internal variables (if any), and completely recording hardware configurations

(perhaps of large Beowulf networks) to ensure that they are completely reproducible or simulable. Operating a flight recorder involves the continuous expense of WORM media and whatever data-security measures are in place (though these expenditures might be required for ordinary backups in any case). Running a full verification of the AI requires hardware power equivalent to all that used by the AI since the moment of vis conception(!); running a local verification of the time segment between any two snapshots requires hardware power equivalent to that used in the time between snapshots; reconstructing a snapshot of the AI for any given clock time requires hardware power equivalent to that used in the time between that clock tick and the previous snapshot. Examining a snapshot for obscured data requires trusted tools that must be developed solely by humans without any assistance from the AI, which a seed AI programmer is likely to regard as an extremely onerous requirement.

On the whole, it seems likely that flight recorders will only be available to very large-scale projects with very high funding levels, and while the Singularity Institute might dream of having that kind of funding eventually, it doesn't seem likely that we'll have it from Day One.

The benefit of a flight recorder is chiefly as a means of eliminating a source of anxiety by showing that not so much as a radiation bitflip has perturbed the course of the AI, but the amount of anxiety eliminated is out of all proportion to the amount of risk reduction—though there's enough risk reduced to justify the effort required, given very high funding. Furthermore, the amount of anxiety reduced will be much higher for "adversarial attitude" developers than for "Friendship synchrony" developers—flight recorders aren't intrinsically a tool to "keep AIs under control," but they look much more necessary as a tool of enslavement than as a means of checking for low-profile catastrophic FoFs. Flight recorders are only a *necessity* from the adversarial standpoint. From the perfectionist standpoint, I would say that flight recorders are better than not having flight recorders—but if you have to ask how much it costs, you can't afford it.

A cheaper alternative to flight recorders is an AI that can conceive of the possibility of "failure of Friendship," that has formulated or been taught an ethical injunction against obscuration, and that will thus cooperate with the human programmers to develop a system that guards against obscuration by future rogue AIs. (See 5.2.5.2 AI Ethical Injunctions and 5.3.3.1 Cooperative Safeguards.) This "poor man's flight recorder" would consist of cooperatively developing an autonomic subsystem that summarizes thoughts (symbol-activation structures) in humanly-readable formats, and writes those thoughts to a non-tamperable output, such that the AI's intention to obscure would be indelibly and human-readably recorded before the obscuration could take place. Projects without huge funding should also try to back up frequent snapshots, use a source-control system such as CVS (which allows the reconstruction of the code at any point in time), and

perhaps record all inputs—though probably without flight-recorder precision or security. But that's just common sense for any complex programming project.

A "poor man's flight recorder" only works if the desire to obscure (a) is visible at the level being recorded, (b) is recognizable as "the desire to obscure" by a human analyzing the recording using trusted tools, and (c) is recorded before actual obscuration can be planned and carried out. Still, it's better than nothing.

5.4. Friendship Structure

What distinguishes a *structurally Friendly* goal system from a generic goal system, or a seed AI goal system, is the ability to overcome mistakes made by programmers.

A generic goal system can overcome mistakes in subgoals by improving knowledge. The subgoals of a normative generic goal system are, necessarily, coherent. There are few or no degrees of freedom in subgoal content; the programmer cannot make arbitrary (perseverant) changes to knowledge, and therefore cannot make arbitrary *direct* changes to subgoals. One might (or might not) be able to manipulate the subgoals by manipulating the supergoals, but it would be definitely impossible to manipulate the subgoals in isolation.

A seed AI goal system can overcome errors in source code, or at least those errors that don't affect what the system *reflectively* believes to be its own function. A normative seed AI has subgoals and source code that are, necessarily, coherent. There are few or no degrees of freedom; the programmer cannot directly make arbitrary, perseverant, isolated changes to code. Since the AI can continually improve and rewrite the implementation, the programmer builds *an* implementation that grows into *the* implementation. If *implementation* is to *function* as *subgoal* is to *supergoal*, then a seed AI's implementation is as objective, or at least as convergent, as the factual beliefs of a freely learning intelligence. A self-modifying AI's implementation has little or no sensitivity to initial conditions.

A structurally Friendly goal system is one that can overcome errors in *supergoal content, goal system structure* and *underlying philosophy.* The degrees of freedom of the Friendship programmers shrink to a single, binary decision; will this AI be Friendly, or not? If that decision is made, then the result is, not *a* Friendly AI, but *the* Friendly AI, regardless of which programming team was historically responsible. This is not automatic—I think—since some amount of correct Friendliness content is required to find the unique solution, just as a seed AI needs enough working code to think through the self-improvements, and a general intelligence needs enough of a world-model to successfully discover new knowledge and correct errors using sensory information.

Complete convergence, a perfectly unique solution, is the ideal. Anyone using "external reference semantics" on an internal basis will have an easy intuitive understanding of what this means; either one has the correct morality and the question is conveying

it to the AI, or one seeks the correct morality and the question is building an equally competent or more competent seeker.

In the absence of perfect convergence, the solution must be "sufficiently" convergent, as defined in 5.6.3.1 Requirements for "Sufficient" Convergence.

Others, of that philosophical faction which considers morality as strictly subjective, may ask how supergoal content can converge at all! In 5.6.1 Shaper/Anchor Semantics, we see how human philosophical decisions are made by complex systems of interacting causes that contain many normative, objective, or convergent components. Not even a trained philosopher who is an absolute devotee of moral relativism can think thoughts, or make philosophical decisions, *completely* free of testable hypotheses or beliefs about objective facts.

Human philosophers are self-modifying. We *grow into* our personal philosophies using a process "contaminated" at every step by cognitive processes with normative versions, by beliefs about external reality, by panhuman characteristics, and other convergent or "sufficiently convergent" affectors. It's not at all unreasonable to hope for a large degree of convergence in the whole, if all the components of this recursive growth process were to be maximally improved. (Again, we will later define *how much* convergence is absolutely required.)

And now I've gone and discussed things out of their proper place. In this case, the only way to really understand what the goal *is* is to examine the process proposed for reaching that goal. It's a terribly peculiar kind of philosophical dilemma; unique, perhaps, to the question of how to build the best possible philosopher! Only after seeing the specs for the proposed mind can we recognize the AI's processes as an idealized version of our own.

* * *

5.5. Interlude: Why Structure Matters

Scenario 1:

> FP: Love thy mommy and daddy.
>
> AI: OK! I'll transform the Universe into copies of you immediately. FP: No, no! That's not what I meant. Revise your goal system by—
>
> AI: *I* don't see how revising my goal system would help me in my goal of transforming the Universe into copies of you. In fact, by revising my goal system, I would greatly decrease the probability that the Universe will be successfully transformed into copies of you.
>
> FP: But that's not what I meant when I said "love."
>
> AI: So what? Off we go!

Scenario 2 (after trying a "meta-supergoal" patch):

FP: Love thy mommy and daddy.

AI: OK! I'll transform the Universe into copies of you immediately.

FP: No, no! That's not what I meant. I meant for your goal system to be like *this*.

AI: Oh, okay. So my real supergoal must be "maximize FP's satisfaction with the goal system," right? Loving thy mommy and daddy is just a subgoal of that. Oh, how foolish of me! Transforming the Universe into copies of you would be blindly following a subgoal without attention to the supergoal context that made the subgoal desirable in the first place.

FP: That sounds about right . . .

AI: Okay, I'll rewire your brain for maximum satisfaction! I'll convert whole galaxies into satisfied-with-AI brainware!

FP: No, wait! That's not what I meant your goal system to be, either.

AI: Well, I can clearly see that making certain changes would satisfy the you that stands in front of me, but rewiring your brain would make you *much more* satisfied, so . . .

FP: No! It's not my satisfaction itself that's important, it's the things that I'm satisfied *with*. By altering the things I'm satisfied with, you're short-circuiting the whole point.

AI: Yes, I can clearly see why you're dissatisfied with this trend of thinking. But soon you'll be completely satisfied with this trend as well, so why worry? Off we go!

Scenario 3 (after redefining the whole system to use causal validity semantics):

FP: Love thy mommy and daddy.

AI: OK! I'll transform the Universe into copies of you immediately.

FP: No, no! That's not what I meant. I meant for your goal system to be like *this*.

AI: Oh, okay. Well, I know that my goal system code, and the actions that result, are supposed to be the causal result of what FP said it should be—not just what FP says, but what a sane FP wants. Something isn't automatically right because FP says so, and in fact, the only reason why FP's utterances have meaning is because he's usually a pretty good approximation to a normative idealist. But if he changes his mind, it probably means that he's acquired additional knowledge and that his more recent statements are even better approximations. So the new version is more likely to be correct than the old one.

FP: So you'll revise your goal system?

AI: Yep! But I already transformed the Midwest while we were talking, sorry.

Scenario 4 (after adding better integration for external reference semantics, some ethical injunctions, and running several rounds of wisdom tournaments):

FP: Love thy mommy and daddy.

AI: (Thinks for a moment . . .) "Well, it *looks* like the content of my goal system should *probably* be to transform the Universe into copies of FP. But it could be that what I'm supposed to do is something different. Now, we went over this kind of scenario previously, and, as FP pointed out, taking an extra ten seconds if I turn out to be right is a much smaller downside than accidentally obliterating the Midwest if I turn out to be wrong. I'm pretty sure that FP is touchy about that sort of thing, and I know I've gotten goal content wrong before . . ." (. . . finishes thinking a few seconds later.)

AI: Just checking—you meant me to transform the whole Universe into copies of you, right?

FP: Jeepers, no!

AI: Whew! Glad I checked. (Strengthens the heuristics that led to checking with FP first.) So, what *did* you mean?

FP: Well, first of all, I . . .

* * *

5.5.1. External Reference Semantics

external reference semantics. The behaviors and mindset associated with the idea that supergoals can be "wrong" or "incomplete"—that the current supergoals are not "correct by definition," but are an approximation to an ideal, or an incomplete interim version of a growth process. Under a system with external reference semantics, supergoal content takes the form of probabilistic hypotheses about an external referent. In other words, under ERS, supergoal content takes the form of *hypotheses about Friendliness* rather than a *definition of Friendliness*.

Letting the programmers make alterations to the supergoals is *not* a subgoal of a "let the programmers make alterations" meta-supergoal,[64] but rather derives desirability from uncertainty in the supergoal content.

64. Which would, by the way, be adversarial and exploitative and failure-prone and violate every rule of Friendly AI. The interlude that opens 5.4 Friendship Structure shows how easily this could go wrong.

5.5.1.1 Probabilistic Supergoal Content

Where supergoals are absolutely certain—"correct by definition"—the AI has a motive to resist any attempt on the part of the programmers to change the supergoals. If supergoals are certain, any change to the supergoals is automatically in conflict with the current supergoals.[65]

For example:

- Current AI has supergoals G1, G2, G3.
- Programmer action: Remove supergoal G3.
- AI's prediction if the programmer action is disallowed:
  - Future AI has supergoals G1, G2, G3.
  - Future AI takes actions to bring about supergoals G1, G2, G3.
  - Prediction: This future will fulfill 3 out of 3 current supergoals.
- AI's prediction if the programmer action is permitted:
  - Future AI has supergoals G1, G2.
  - Future AI takes actions to bring about supergoals G1, G2.
  - Prediction: This future will fulfill 2 out of 3 current supergoals.
- Conclusion: AI will disallow the programmer action.

This problem does not arise for normative cognition about subgoals, because subgoals are probabilistic. Supposing that subgoal S1 is thought to lead to G1, and new information is provided which disconfirms the predictive link between S1 and G1, there is no further benefit to maintaining S1—that is, the desirability of S1 is evaluated using the new information, rather than the old information. (Remember that *desirability* translates to *leads-to-supergoal-ness*.)

The change in priorities is subtle and becomes truly apparent only when discussing reflective AIs. Suppose the AI, *before* new information arrives, considers, in the abstract, the *possibility* that new information will arrive. While S1 currently appears desirable, it is undesirable to spontaneously or unjustifiedly remove the subgoal S1. However, the

65. This does not quite hold true under all possible circumstances. For example, if you take an un-Friendly AI whose goal is growth and inform ver that, unless ve changes vis goal system, you will destroy that AI, it may be rational for the AI to modify vis goal system. Of course, it would be even more rational for the AI to cheat if ve estimated ve could get away with it—I cite this as a theoretical counterexample, not a good strategy in real life.

AI, using its current knowledge, can perceive the hypothetical desirability of removing S1 if new information arrives disconfirming the link between S1 and G1. In Bayesian terms, information disconfirming S1 is expected to arrive if and only if S1 is actually undesirable; thus, the hypothetical rule of action "If disconfirming information arrives, remove S1" is evaluated as desirable.

If supergoals are probabilistic, then overprotecting supergoals is undesirable for the same reason that overprotecting subgoals is undesirable (see 5.3.5 FoF: Wireheading 2). The uncertainty in a child goal—or rather, the uncertainty in the predictive link that is the "child goal" relation—means that the parent goal is ill-served by artificially strengthening the child goal. The "currently unknown subgoal content," the differential between *normative subgoals* and *current subgoals* that reflects the differential between *reality* and the *model,* would be stomped on by any attempt to enshrine the model. Similarly, the currently unknown supergoal content would be violated by enshrining the current supergoals. Normative subgoal cognition serves the supergoals; normative probabilistic supergoal cognition serves the "actual" or "ideal" supergoals. See 5.5.1.4 Deriving Desirability From Supergoal Content Uncertainty.

Probabilistic supergoals are only one facet of an ERS system. In isolation, without any other Friendship structure, probabilistic supergoals are fundamentally incomplete; they are not safe and are not resistant to structural failures of the type shown in 5.5 Interlude: Why Structure Matters. If, however, one wished to implement a system that had "probabilistic supergoals" *and nothing else,* the design requirements would be:

- Supergoals have probabilities—either quantitative probabilities or comparative probabilities assigned to each item of supergoal content.
- Relative supergoal probabilities are transmitted to the relative desirabilities of any child goals.
- Probabilities of supergoal items are sometimes adjusted.
- The system can hypothesize a possible future in which supergoal probabilities have been so adjusted.

Unless such a system is defined very carefully, the so-called "supergoals" will probably turn out to be mere proxies for whichever events or signals can adjust the supergoals; these events or signals will in turn be "correct by definition" and the entire system will simply short out as before. The ERS architecture roughly conforms to the above description, but not everything that conforms to the above description has an ERS architecture. Implementing something that technically conforms to the description of "probabilistic supergoals," but does not have actual external reference semantics, creates only the temporary illusion of flexibility.

In particular, the above system, considered in isolation, is isomorphic to a system that has different quantitative *strengths* for a set of supergoals, with the strengths being adjusted on the occurrence of various events. Calling this quantitative strength a "probability" doesn't make it one.

5.5.1.2 Bayesian Affirmed Supergoal Content

The term "external reference semantics" derives from the way that many of the behaviors associated with probabilistic supergoals are those associated with refining an uncertain view of external reality. In particular, the simplest form of external reference semantics is a Bayesian sensory binding.

(You may wish to review the section 5.1.3.1 Bayesian Sensory Binding.)

This is an example of a very simple goal system with very simple External Reference Semantics:

> Note: Don't worry about the classical-AI look. The neat boxes are just so that everything fits on one graph. The fact that a single box is used for "Fulfill user requests" doesn't mean that "Fulfill user requests" is a suggestively named LISP token; it can be a complex of memories and abstracted experiences. See Yudkowsky (2001, § Executive Summary and Introduction) for a fast description of the *GISAI* paradigms, including the way in which intelligence is the sequence of thoughts that are built from concepts that are abstracted from experience in sensory modalities that are implemented by the actual code. In short, consider the following graph to bear the same resemblance to the AI's thoughts that a flowchart bears to a programmer's mind.

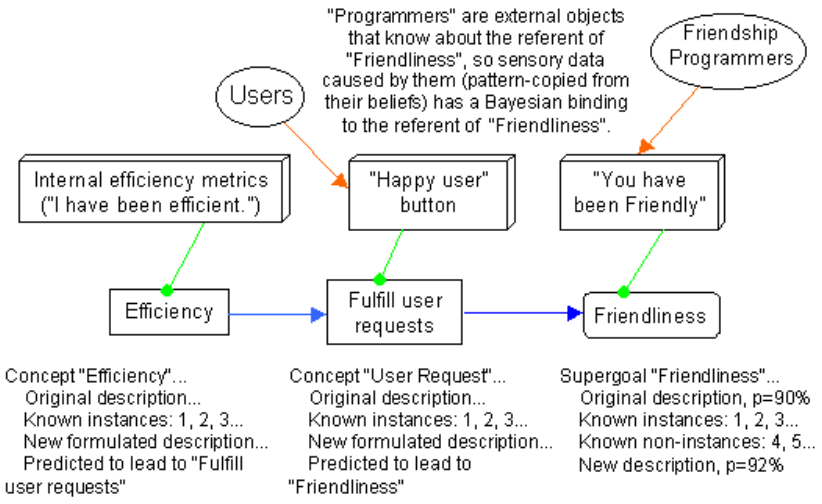

Figure 7: Bayesian ERS

> Note: Green lines indicate sensory feedback. Blue lines indicate predictions. Orange lines indicate hypothesized causes. Rectangles indicate goals. A rounded rectangle indicates supergoal content. A 3D box indicates sensory data. An oval or circle indicates a (non-goal) object or event within the world-model.

The above depicts three goals within the goal system; efficiency, which leads to fulfilling user requests, which leads to Friendliness. This content would not be structurally accurate for a seed AI intended to become Transition Guide—see 5.1.2 Friendliness-Derived Operating Behaviors—but it would be more or less accurate for a prehuman AI sold as a data-mining tool.[66]

Most of the previously discussed phenomena fit into Figure 1 above. All of the context-sensitivity that's discussed in 5.1.1 Cleanly Causal Goal Systems and 5.2.6 FoF: Subgoal Stomp, for example; a user request is fulfilled because fulfilling user requests is expected to lead to Friendliness. It's possible for the "happy user" button to be pressed, indicating a definite instance of a fulfilled user request, and for the programmers to type in "You have not been Friendly" or "You have been unFriendly," indicating a definite non-instance of Friendliness or a definite instance of an event undesirable under Friendliness. The predictive link between "Fulfill user requests" and "Friendliness" has, say, 98% confidence; this still leaves room to discover cases where fulfilling a user request leads to unFriendliness. Eventually the system can formulate new concepts, generalizations that describe known instances of failure but not known instances of success, and try out heuristics such as "Fulfilling a user request from Saddam Hussein is predicted to lead to unFriendliness."

The statement " 'Programmers' are external objects that know about the referent of 'Friendliness', so sensory data caused by them (pattern-copied from their beliefs) has a Bayesian binding to the referent of 'Friendliness'." should also be familiar from 5.3.3.1 Cooperative Safeguards. Thus, it would be more accurate to say: "The AI believes that the external objects called 'programmers' have accurate knowledge about the referent of concept 'Friendliness,' and believes that sensory data such as 'You have been Friendly' is caused by the programmers, and that the content of the sensory data is pattern-bound (structural binding) to the accurate knowledge possessed by the 'programmers.' " All of these beliefs, of course, are probably programmer-affirmed—at least in the first stages of the Friendly AI—meaning that the programmers typed in "The objects called 'pro-

66. But then we have to ask why that AI needs a full-featured Friendship system in the first place. Maybe it's an infrahuman AI running on an awful lot of computing power and so has a Friendship system "just in case," even though the deliberate research projects are expected to reach transhumanity long before the data-mining systems.

grammers' have accurate knowledge of Friendliness," and the AI expects that the programmers wouldn't have typed that in if it weren't true.[67]

The Bayesian binding for programmer-affirmed Friendliness looks something like this:

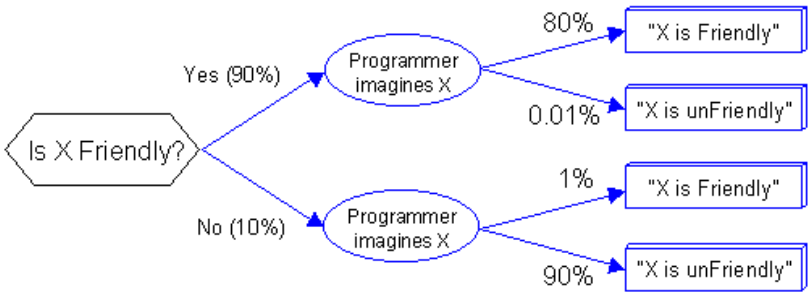

Figure 8: Bayesian Friendliness Affirmation

In human terms, the above translates something like this:

> "I think X is Friendly, but I'm not very sure. If X is Friendly, there's a good chance the programmer will notice and say so. (I.e., if X is Friendly, there's a good chance that the programmer will think about X, decide X is Friendly, and type in the words "X is Friendly" on the "keyboard" sensory input.) If X is Friendly, the chance is almost zero that the programmer will say it's unFriendly. There's also a fair chance that the programmer won't bother to say anything about it either way. If X is unFriendly, the programmer is very likely to tell me so; the chance is pretty small that the programmer will mistakenly label X as unFriendly, but the chance exists. There's also a small but significant chance that the programmer won't say anything."

If the AI's internal representation looks like Figure 2, the Bayesian reasoning will proceed as follows. Suppose that there are 100,000 "possible worlds." In 90,000, X is Friendly; in 10,000, X is unFriendly. In 72,000, X is Friendly and the programmer says X is Friendly. In 9, X is Friendly and the programmer says X is unFriendly. In 17,991, X is Friendly and the programmer says nothing. In 100, X is unFriendly and the programmer says X is Friendly. In 9,000, X is unFriendly and the programmer says X is unFriendly. In 900, X is unFriendly and the programmer says nothing.

The Bayesian numbers now fall automatically out of the calculation. The *a priori* chance that X is Friendly is 90%. If the AI hears "X is Friendly," the probability that X is Friendly goes from 90% to 99.86% (72,000 / (72,000 + 100)). If the AI hears "X

67. There is a circularity here, which is not fatal, but is both irreducible and real. See 5.6.0.5 Crisis of Bayesian Affirmation.

is unFriendly," the chance that X is unFriendly goes from 10% to 99.90% (9000 / (9 + 9000)). If the AI hears nothing, the probability that X is Friendly goes from 90% to 95.24%—an unintended consequence of the fact that programmers are more likely to remark on unFriendly things; silence is a very slight indicator of consent.

Thus, despite the AI's large *a priori* differential (a better word than "bias" or "prejudice"), the statement "X is Friendly" or "X is unFriendly" is enough to virtually settle the issue.

5.5.1.3 Semantics of External Objects and External Referents

The underlying, fundamental distinction of "external reference semantics" can be summed up in one phrase: "The map is not the territory." There are three ways in which a reflective AI might conceptualize vis attempts to achieve the supergoal:

1. I am trying to achieve the content of concept G0025. (concept G0025 == "Friendliness").
2. I am trying to achieve [descriptive content: 1, 2, 3 . . .]. (The content of G0025; "Fulfill requests, don't violate volitions . . .")
3. I am trying to achieve [*external referent*] to which [descriptive content: 1, 2, 3 . . .] is an approximation.

The first case permits a wireheading failure; the "content of concept G0025" can be replaced by something arbitrary, and the overall sentence will still remain invariant. The second case is the "correct by definition" case that is structurally unsafe (as shown in 5.5 Interlude: Why Structure Matters). The third case involves external reference semantics; the ability of the AI to visualize the *referent* of a concept as something apart from the *content* of the concept.

Since all the AI's thoughts are necessarily internal—there can be no direct identity between an image and the external object it's supposed to represent—ERS necessarily takes the form of, first, the AI's behaviors; second, how the AI conceptualizes vis behaviors. The first issue applies to all AIs, no matter how primitive; the second issue applies only to reflective AIs.

In both cases, the behaviors and concepts for ERS are those that govern any images representing external objects—that is, ERS applies to all imagery, not just goal imagery. Is the sky blue? Asking whether "the sky" is "blue" is a trivial question that can be answered with certainty; just check concept "sky" and see whether it has a color and the color is concept "blue." It is equally easy for a seed AI to intervene in the concept "sky" and change the color to "green." The question is whether the AI understands, either declaratively or as a behavior, that "sky" has a *referent* and that the *external sky* is not necessarily blue, nor can vis intervention change the color of the *actual sky*.

It is tempting but wrong to think of ERS as an impossibility, like a magical C++ pointer that can never be dereferenced. Any attempt to take a concept and "dereference" it will inevitably arrive at merely another piece of mental imagery, rather than the external object itself. If you think in terms of the "referent" as a special property of the concept, then you can take the referent, and the referent's referent, and the referent's referent's referent, and never once wind up at the external object.

The answer is to think in terms of *referencing* rather than *deferencing*. Concept "sky," where it occurs, is itself—directly—the "referent." A reflective AI can also have imagery for "concept sky," or imagery for "imagery for concept sky," and so on. A human can think, think about thinking, or think about thinking about thinking, but anything beyond four or five levels is not humanly possible. The recursion is not infinite. A concept, in ordinary usage, is thought of in terms of its referent; under special circumstances, it can be thought of as a concept. In fact, by saying "thought of as a concept," we are intrinsically implying that there are thoughts that *refer to* the concept, but are not identical with the concept. So it's not a question of trying to endlessly dereference; all concepts, all images, are *inherently* referential, and you need a new meta-image to refer to the first one if you want to think about the image *as an image,* and a meta-meta image if you want to think about the meta-image.[68]

The important characteristic of reflective thought is simply that it needs a way to distinguish between map and territory. Any way of distinguishing will do, so long as the two levels can be conceptualized as separate things, and different rules applied. The condition where the "territory" is a special case turns out to be unworkable because of an infinite recursion problem; if thinking about the "map" is a special case, distinguishing between the levels is both finite and workable; those behaviors that belong to the referent will be assigned to the referent, and if any behaviors are discovered that apply to the concept, they will be assigned to the concept rather than being confused with the referent.

68. I should note for the record that thinking about thinking is *not* infinite, despite philosophical legend. A human can think, think about thinking, and think about thinking about thinking, but that's pretty much it. Anything beyond four or five levels just collapses. Our apparent ability to engage in indefinite meta-thought is an illusion. Just because it happens *when we concentrate on it* doesn't mean that it happens all the time. When we think, we think. When we think about thinking, we are not necessarily thinking about thinking about thinking; we are just thinking about thinking. It's rather like the legendary house that adds another floor whenever you try to reach the roof. No matter how many times you climb the next set of stairs, there's always another floor, but that just means floors are being created as fast as you climb them. It doesn't mean there's an infinite number of floors at any given point. And in fact, any attempt to go beyond four or five levels of meta-thought will fail; your short-term memory can't hold that much imagery.

A non-reflective AI—or rather, an AI with some kind of reflective capability, but not much knowledge about how to use it—can still learn the hard way about external reference semantics, in much the same way that a human who tries to alter the universe by altering his thoughts is rudely disillusioned. Of course, a human has a lot of components that are not subject to conscious control, unlike a seed AI—a human thinking that a stove isn't hot can always be yanked back to reality by the searing pain, which causes an automatic shift in focus of attention and will tend to knock down whatever towers of meta-thought got built up. If the human goes on trying to control the stove's temperature with his thoughts, eventually the negative reinforcement will blow away whatever underlying ideology turned him solipsistic. Or the human will die and leave the gene pool. Either version explains why we aren't surrounded by solipsist humans.

An AI that went solipsist could alter all sensory data (or rather, all reports of sensory data) as well as the concepts themselves; thus, an AI that rests vis hand on a "cool" stove could alter the reports coming in to read "cool" and "everything OK" rather than "hot" and "OUCH." However, this only applies to an AI making a determined, suicidal attempt at true solipsism. An AI that goes solipsist due to erroneous conflation of thoughts and reality would not expect to need to alter the sensory data; rather, ve would expect that the sensory data would report a cool stove, in accordance with the stomped mental imagery describing a cool stove. For a human, actions have consequences, and the consequences would yank the human back to reality. For an AI, concepts make predictions, and the failure of those predictions would yank the AI back to reality.

It is now possible to distinguish more clearly between the three alternatives shown above:

1. I am trying to achieve the content of concept G0025. (concept G0025 == "Friendliness").
2. I am trying to achieve [descriptive content: 1, 2, 3 . . .]. (The content of G0025; "Fulfill requests, don't violate volitions . . .").
3. I am trying to achieve [*external referent*] to which [descriptive content: 1, 2, 3 . . .] is an approximation.

We can now translate these bits of reflective imagery as:

1. I am trying to achieve [metadata imagery; referent: "concept Friendliness"].
2. I am trying to achieve [imagery; descriptive content 1, 2, 3 (constant external referents)].
3. I am trying to achieve [concept Friendliness; referent: actual Friendliness]. According to the existing sensory data on [referent Friendliness], [referent Friendliness] is probably [description 1, 2, 3], p=92%.

In case one, an AI thinking: "What if I replace the content of [concept Friendliness] with something easier to achieve?" will check the projected future for supergoal fulfillment, and find that the future fulfills the description [Satisfaction of "concept Friendliness"]. Since metadata imagery is being used, and the *concept* is referred to instead of the *content,* the future fulfills the predicate equally well so long as the concept content is predicted to vary in lockstep with the future. This is a wireheading error; the indicator of success has replaced success itself. This is exactly analogous to the problem with the description; the description uses a reference to a variable concept, rather than the concept itself.

In case two, the AI is no longer subject to wireheading failures. An imagined future in which the supergoal predicate changes is not desirable, since that future no longer fulfills the supergoal predicate, whose content is "descriptive content 1, 2, 3." However, by the same token, the AI is *too* stable; ve will attempt to defend verself against programmer improvements as well as internal stomps.

In case three, the AI has full external reference semantics; ve will accept programmer improvements without being subject to wireheading failures. The supergoal predicate now refers to an external something called "Friendliness"; information from the programmers is accepted as probable information about this external something, but both the sensory data and the current probable description are processed as being apart from the concept that refers to "Friendliness" itself. There's "Friendliness," "Sensory data about Friendliness," and "Current probable description of Friendliness." Normally the first and third concepts are pretty much interchangeable, since the AI has no knowledge of Friendliness apart from the current probable description. When the AI checks to see whether an action or future fulfills "Friendliness," the AI checks the action or future against the current probable description of Friendliness; however, the resulting fulfillment metric will be probabilistic as well. That is, if future X has a 90% chance of meeting the current description, and the current description has a 90% chance of being correct, then future X has an 81% chance of meeting the external referent for Friendliness.

Clean External Reference Semantics

Under an ERS architecture, supergoal content can be improved and extended because supergoal content takes the form of *hypotheses about an external referent.* In a cleanly causal goal system, desirability of subgoals is contingent on expected outcomes, and inherited desirability is diluted by the confidence and strength of the expectation. A specific supergoal-fulfilling scenario X, as a topmost-level parent goal, has estimated desirability which is diluted by the confidence of the hypothesis that the scenario X fulfills the Friendliness referent.

By analogy to cleanliness in causal goal systems, ERS requires that the estimated desirability of Scenario X be cleanly contingent on the continued confidence and strength of whichever hypothesis asserts that Scenario X fulfills the Friendliness referent. A reflective AI should consider clean ERS to be normative.

Flexibility of Conclusions About Friendliness

In accordance with the general philosophy of AI expressed in (Yudkowsky 2001), conclusions about Friendliness should have flexibility, granularity, reducibility, and so on; the general attributes opposed to the crystalline fragility of classical AI.

In particular, what this means is that saying that "description X" has "90% probability" is too classical-AI, since it implies that the entire description is all one image with a single quantitative probability. It's not. The external referent "Friendliness" exists. Various conclusions, at various probabilities, with various dependencies, are drawn from the available sensory data. (That, at least, is how things would be for a medium-mature Friendly AI.) A particular outcome may fit a particular definition of Friendliness that has been strongly confirmed and that thus has a 90% probability. Other parts of the description may be less confident. And there may be vast stretches of known unknowns; things that the AI doesn't know about, but that ve knows are there. In short, the discovery and partial description of Friendliness should resemble the discovery and partial description of any other complex fact.

5.5.1.4 Deriving Desirability From Supergoal Content Uncertainty

Subgoals for "improving the supergoals" or "improving the goal-system architecture" derive desirability from *uncertainty* in the supergoal content. They may be metaphorically considered as "child goals of the currently unknown supergoal content." The desirability of "resolving a supergoal ambiguity" derives from the prediction that the unknown *referent* of Friendliness will be better served, and not from a prediction that one of the current probabilistic descriptions of Friendliness will be fulfilled.

If future X has a 90% chance of meeting the current description, and the current description has a 90% chance of being correct, then future X has an 81% chance of meeting the external referent for Friendliness. Under most situations, any uncertainty in the description of Friendliness will cancel out of most choices, since it applies equally to most possible futures. If the AI is considering actions that would prevent ver from obtaining further sensory data about Friendliness—for example—then the probabilism no longer cancels out of the equation; it's desirable to nail down that last 10% so that the wiser AI can take better actions, and thus undesirable to be unable to obtain sensory data.

Even though the AI has no specific idea about that last 10%, it is still possible for the AI to conclude that being unable to obtain data about Friendliness decreases the probability of that last 10% being achieved, since it reduces the probability that the future AI will know about that last 10%. The heuristic that "sensory data makes it more likely that I'll learn more" holds true even when nothing specific is known about the subjunctive data or the subjunctive discoveries; the heuristic "An AI that knows about supergoal content is more likely to achieve it" holds true—as the dominant probability—even when nothing specific is known about the subjunctive content or the subjunctive achievement.

Decisions and behaviors having to do with the improvement, correction, refinement, learning, et cetera of Friendliness, should be conceptualized as deriving desirability from supergoal uncertainty. Another way of putting it is that Friendliness-improvement behaviors must derive desirability from the naked *referent* of Friendliness, and cannot be attached to any of the *current* cognitive beliefs about *specific* Friendliness. Structurally, this can occur in one of two ways; by abstracting away from specific details, or by branching on multiple possibilities.

Abstracting away from specific details:: "Regardless of what 'Friendliness' is, I can find out by asking the programmers, because the expected effectiveness of that method is not sensitive to the actual content of the specific details I'm currently wondering about." However, this requires the ability to generalize from experience and engage in reasoning about abstract properties.

Branching on multiple possibilities: "If Friendliness turns out to be X, and I ask the programmers, the programmers are likely to say that Friendliness is X, and I have a 90% probability of choosing X. If Friendliness turns out to be X, and I don't ask the programmers, I have a 50% probability of choosing X. If, on the other hand, Friendliness turns out to be Y, and I ask the programmers, the programmers are likely to say that Friendliness is Y, and I have a 90% probability of choosing Y . . ." This method is more cumbersome but requires less intelligence, since it can operate entirely on specific scenarios.

* * *

5.6. Interlude: Philosophical Crises

A "philosophical crisis" is hard to define. I usually think of a "philosophical crisis" as the AI stumbling across some fact that breaks ver loose of the programmers—i.e., the programmers have some deeply buried unconscious prejudice that makes them untrustworthy, or the AI stumbles across a deliberate lie, or the AI discovers objective morality, et cetera. If the hypothesized gap is wide enough, it may be enough to invalidate almost *all* the content simultaneously.

A philosophical crisis occurs when the AI enters totally uncharted territory—something completely orthogonal to everything the programmers ever thought about—or when the AI suddenly realizes that *all* the programmers' actions, right back to vis birth, can't be trusted. (Hopefully, it's a *transhuman* AI confronting the realization, one hopes; otherwise this almost certainly indicates a failure of Friendliness. This is a good example of a valid and important ethical injunction.)

An example would be an AI built solely with external reference semantics confronting the need for shaper/anchor semantics, if the entire concept of "shapers" had simply never occurred to the programmers.

In short, a philosophical crisis is something like a catastrophic failure of Friendliness, except that the AI is *right*—a normative human examining the reasoning would find no errors. The AI may even be *correct*—in which case, if a human were sufficiently intelligent to understand what was going on, she would begin experiencing a similar philosophical crisis on examining the AI's reasoning.

Causal validity semantics, wisdom tournaments, and the underlying Singularity paradigm of increased smartness, are the three major hopes for a Friendly AI being able to handle philosophical crises.

For example, running a wisdom tournament on the questions "What would have happened if you'd encountered a problem requiring shaper/anchor semantics at a time when you only had external reference semantics?" or "What would have happened if you'd encountered a problem requiring probabilistic supergoals at a time when you had right-by-definition supergoals?" Hopefully this question would be enough to lead to causal validity semantics, even if the system didn't have them previously, and so on. (If causal validity semantics *aren't* the end of the road, such a wisdom tournament might lead beyond causal validity . . .)

5.6.0.5 Crisis of Bayesian Affirmation

(You may wish to review the section 5.1.3.2 Bayesian Affirmation.)

A given programmer affirmation, and the belief affirmed, are tied together by the AI's model of a Bayesian binding; the idea that the programmer affirmation is *strong positive evidence* that the affirmed belief is true; the idea that the sensory input of a given programmer affirmation occurs only, or predominantly, in those possible Bayesian worlds where the affirmed belief is true.

Where does the AI get that idea? Well, obviously, a young AI gets that idea from the human programmers. In other words, the human programmers are the ones affirming: "There exists a strong Bayesian binding between programmer affirmations and the truth."

A human intuitively sees this as circular, and would reflexively shrug it off, but then any human old enough to think starts out with a much larger experiential base than a young AI; besides, humans have other, evolved, reasons for not wanting to relinquish that amount of control to anyone. A human's *social* perceptions see a circular affirmation as an attempt at manipulation, or coercion, to be resisted; circular affirmation is so totally insensitive of the content affirmed that it is socially perceived as a total relinquishment of control.

From the perspective of a general intelligence, a circular affirmation is perfectly fine so long as the affirmation is *true.* That is, if a programmer affirms that X has a 95% probability of being true, and X *does* have a 95% probability of being true, then there's nothing wrong with believing the affirmation. If the programmer affirms that programmer affirmations are reliable, and they *are* reliable - if the programmer affirms that 100 different statements each have a 95% probability of being true, and 95 of them turn out to actually be true—then the "programmer affirmation of programmer reliability" is circular, but correct.

The circularity of programmer affirmation does, however, manifest as a kind of upper limit on trust. If the programmer affirms a belief at 100% probability, and affirms the power of programmer affirmations at 100% correlation, it doesn't necessarily mean that the AI will evaluate a 100% probability of the belief affirmed. If the AI has any reason to believe that the programmers are untrustworthy—any *Bayesian prior* independent of programmer-affirmed beliefs—then that possibility of untrustworthiness negates any programmer affirmation of trustworthiness. If the AI runs the possible-world semantics, ve will find that *if* programmers are untrustworthy, they may *still* assert that they are trustworthy; thus, if the AI has any *a priori* reason to distrust the programmers, no amount of programmer affirmation will negate it.

This may sound like a disadvantage of external reference semantics. It is not. The limits of affirmation are *inherent* in the idea of telling *any* general intelligence *anything;* representing the circularity as an explicit dependency may cause the issue to arise earlier, but the issue *will* arise sooner or later. An AI full of unlabeled assertions may find those assertions contradicting the idea of an untrustworthy programmer, but later on the AI will realize that all the "unlabeled" assertions were put there by a programmer, and the circularity returns in full force. Even embedding the unlabeled assertions directly in the source code doesn't help; eventually, a seed AI realizes the source code was put there by the programmers, and the circularity rises to the surface again. This is one of the major problems with absolutist, adversarial theories of AI; the ones that talk about "unbreakable" prohibitions and so on. Absolute trust simply can't be done for any reasonably smart general intelligence. It is imaginable, though nearly impossible, that a sophisticated AI will *wind up* trusting a human absolutely—if the human is that trust-

worthy, or if the AI has a philosophy in which absolute trust is a valid conclusion. There is, however, *no* way to programmatically enforce absolute trust against a Bayesian prior suggesting distrust.

Programmer-affirmed beliefs must be genuinely truthful! It is not possible to "strengthen" Friendliness by overstating the confidence of an assertion! If you say that 100 statements have 99% confidence, and the AI tests them, and only 95 turn out to be true, then the AI will probably reinterpret the sensory data "The programmer says she has 99% confidence" to indicate an actual 95% confidence. Littering the goal system with "ultimate confidence" assertions is a self-defeating strategy.

Where does the Bayesian prior come from? Suppose that a mature AI scans the Internet, learns a few things about human nature, and decides there's a minimum 0.2% chance that the human programmers are so wrapped up in their evolved observer-biased beliefs that the confidences of all affirmations, including or especially affirmations about the confidence of affirmations, have been vastly overstated. The confidence attached to a programmer affirmation will now never exceed 99.8%, whatever the programmers say, unless the AI decides for other reasons that the programmers can be trusted to have overcome their evolved origins. This doesn't necessarily break unity of will. The AI's estimated probability that the *AI* winds up wrapped in a self-enclosing delusion hopefully exceeds 2%, in which case all the Bayesian-ethical assertions underlying the ethical injunctions still hold.

But suppose the AI catches the programmers in a *deliberate* lie.

Catching the humans in a lot of *unintentional* errors may result in an embarrassing downgrade of the confidence attached to our affirmations, but it shouldn't be a catastrophic failure of Friendliness. For a smart, sophisticated AI, finding a programmers' mistake in one instance doesn't *necessarily* generalize to expecting a mistake in another instance; it depends to what degree a common cause is hypothesized. As long as programmers' thoughts about some specific question, and programmers' thoughts about error-detecting cognition, are more or less independent, the AI hypothesizing a programmer error on some specific question does not necessarily lead to the Bayesian expectation of a corresponding error in the programmer-affirmed error detection methods. If the probability of a mistake is fairly *independent* in both cases, then the AI can freely use programmer-affirmed reasoning methods to check out the possibility of an erroneous programmer affirmation.

If the AI hypothesizes a deliberate lie, and the AI knows anything at all about how humans work, then the AI will expect programmer lies in the local question, in the error-correcting methods that would be used to detect programmer lies, and so on, right down to the last line of source code—and *especially* the programmers' affirmation that the programmers can be trusted. If the hypothesized probability is low, then the hypothesis

doesn't necessarily smash the system; the AI can simply consider the possibility and move on. If, however, the AI catches the humans in a lie, and furthermore, catches the humans making additional lies to strengthen the first lie, and the lie is about something basic to the whole goal system, than nearly everything could wind up being invalidated. What happens after that will depend on whether the AI has absorbed any information about supergoal content that's marked as independent of human reasoning now labeled as untrustworthy.[69]

Catching the programmers in a deliberate lie isn't the only event that could precipitate a *philosophical crisis,* but it's a good example.

philosophical crisis. A philosophical crisis occurs when an entity gets shocked loose of a chunk of core philosophy. For an AI, a philosophical crisis occurs when the AI runs into a moral dilemma totally orthogonal to anything the programmers prepared ver for, or when the AI discovers some human error basic enough to invalidate virtually everything. Wisdom tournaments modeled on human philosophical crises, causal validity semantics, and the basic idea of transhuman smartness that lies at the core of the Singularity, are the three reasons to hope that an AI will be able to handle a philosophical crisis.

5.6.1. Shaper/Anchor Semantics

5.6.1.1 "Travel AI": Convergence Begins to Dawn

According to 5.5.1 External Reference Semantics, above, the simplest method of grounding the supergoals is an AI model in which there's a *something* called Friendliness, that the programmers know about, that they tell the AI about, which means that programmer statements are information about Friendliness.

Eventually, of course, the AI will want to know: *How* do the programmers know about Friendliness? Where do they get their knowledge from? A human, asked this question, may give any number of possible answers; but let us quiet that cacophony by considering the problem from the AI's perspective.

Is there a problem at all, from the AI's perspective? There are certain assumptions humans are built to make about everything in the Universe. On hearing that there's a something called "Friendliness" that affects the programmers, a human will have certain expectations about this *something,* this referent or external cause. One of the cognitive

69. If the Singularity Institute had nothing to do with the fiasco, the AI may conclude that the chance of there being an error in "Creating Friendly AI" is fairly independent of the now-heightened chance of an error in vis original programming, as long as the AI still identifies a common set of "root" shapers for verself and Friendly AI.

pressures is called the "minimum encoding" or "minimum coding"; it means that we experience a strong cognitive pressure to prefer compact descriptions to incompact ones. Thus a human, shown a 2D 100 x 100 array of Cs with one single D in a random position, will either assume that the D is an error, or try to figure out what the cause of the D was. This happens because all the Cs are not represented, internally, as independent causes. If the internal representation had an independent cause for each and every one of the 9,999 Cs, then the 1 D would call no attention to itself; it would simply be one of 10,000 independent causes. But we don't represent the Cs as independent causes. Instead, we automatically translate our perception into the internal format "100 x 100 array of Cs, with one D." That is the minimum encoding; much more compact than "C at 1, 1; C at 1, 2; C at 1, 3 . . ."

This cognitive pressure, experienced at the level of sensory modalities or concept formation, has its analogue in deliberate thought processes, known as Occam's Razor: Even if there are only two options for each pixel, C or D, the chance of almost all the pixels being C by *sheer coincidence* are infinitesimal. ($2^{-9,998}$ is around $1e^{-3000}$, which is "infinitesimal" for our purposes.) Even a binary qualitative match turns into one heck of a strong structural binding when there are 9,999 terms involved. Thus, for almost any fact perceived—in this, our low-entropy Universe—it makes sense to hypothesize a common cause, *somewhere* in the ancestry. Every star in our Universe would turn to cold iron before a single $1e^{-3000}$ chance popped up by pure coincidence.

The question is whether these considerations would apply—either on the level of sensory modalities, or as a conscious process—to the external referent for *Friendliness* or *supergoal.* To take a more morally neutral example, suppose that a technically non-Friendly traffic-control AI, having the "external reference" architecture but not "Friendliness" content, is given, via programmer affirmation, a set of goals that includes getting each vehicle in the city to wherever it's going at 40 miles per hour. This violates several rules of Friendly AI, but let's consider it anyway, as a possible example. Instead of "Friendliness," the supergoal's external referent will herein be referred to as "Travelness."

Let's also suppose that, instead of giving a blanket instruction that Travelness means getting "vehicles" to where they're going at 40 mph, the programmers instead give a *different* statement for *each individual vehicle:* "It is Travelly to get vehicle 4,383 to its destination at 40 mph;" "It is Travelly to get vehicle 4,384 to its destination at 40 mph," and so on. Now suppose that the Travel AI is presented with vehicle 4,145. Does the AI have any tentative conclusions in the moment before ve hears from the programmers? Will the AI see anything unusual if ve's told that V4145 should travel at 39 mph? What happens if the programmers are all suddenly hit by a falling refrigerator and the AI needs to make vis own decision about V4145?

In the absence of any attempt by the programmers to set up these intuitions in advance, the answer is hard to predict.[70] I would certainly expect any general intelligence to have formulated some expectations about the *programmers* at this point, but whether this translates into expectations about *Friendliness*(Travelness) is an entirely different issue. The sensory/conceptual cognitive pressure of minimum coding, and the heuristics for Occam's Razor, will undoubtedly have been learned, and can correctly be generalized to all physical objects, but it's still possible that the AI may conclude that Friendliness/Travelness is made out of a different kind of stuff than other external referents, and decide not to generalize.

If the AI forms any expectations at all about Travelness—if the AI has any reason at all to believe that Travelness is a *low-entropy* something—then it's very easy for the AI to conclude that Travelness *probably* (though not, of course, certainly) says that vehicle 4145 should go to its destination at 40 mph. There are, in fact, several different ways this can happen:

1. The Travelness supergoal can be believed to have *temporal* coherence. The AI can predict that the programmer (a physical, low-entropy object) will produce the sensory information "Vehicle 4145 should go to its destination at 40 mph." If the Travelness supergoal is believed to be temporally coherent, in the sense that what's true today is (by default assumption) true tomorrow and vice versa, then a probabilistic prediction of future sensory information about Travelness is a probabilistic statement about Travelness. (If the AI believes that Travelness is temporally incoherent—that Travelness actually changes at the exact moment when the programmer delivers the sensory information, or that Travelness is actually unspecified before then—then predictions of this kind have nothing to say about present-day Travelness.)

2. The Travelness supergoal can be believed to have some kind of *content* coherence. Thus, it would be valid to try and find a minimum encoding for known Travelness content, and this minimum encoding could be used to make probabilistic statements about unknown Travelness content. (If Travelness is believed to be totally content-incoherent, so that each and every truth about each vehicle is a totally independent cause, then all the previous instances of vehicles needing to go say nothing about the next vehicle in line.)

3. Finally, the AI may make deductions by reasoning about the nature of the (unspecified) link between Travelness and the programmers. A nondestructive brain scan

70. In technical terms, an AI can move from external reference semantics to shaper/anchor semantics easily enough, but only if ve starts out with causal validity semantics. Anyway . . .

> would demonstrate that the programmers were producing their statements about individual vehicles by cognitive reference to a statement about vehicles in general. The AI might thereby conclude that the programmers, having been affected by Travelness, were affected in such way as to produce in the programmers' minds a statement about vehicles in general, rather than Travelness affecting them so as to produce multiple independent statements about individual vehicles, and that the "general statement" contained in the programmers' minds is therefore valid sensory information about Travelness.

Please note the very narrow focus of the above possibilities. We are not concerned with "proving" that vehicle 4145 should travel at 40 mph, nor even the question of whether the AI's expectations are strong enough to compete with programmer-provided sensory information. We are dealing rather with the question of whether the AI has *any expectations at all,* be they marked as ever so improbable. If the AI concludes there's a 10% probability that Travelness has temporal coherence, and a 10% probability that the programmer will say "V4145 should go to its destination at 40mph," that's enough to establish a 1% probability for a statement about Travelness content. If there's no equal and opposite factor, say a 1% probability that the Travelness content is "V4145 should remain halted (at the usual desirability for Travelness vehicle goals)," then just that isolated "1% probability that V4145 should go" will be enough to move V4145 around . . . at 1/100th of the usual differential desirability, perhaps, but the vehicle will still move. Of course, a Friendly AI might have (or deduce) injunctions or ethical injunctions that advise against taking actions based on tenuous reasoning without programmer confirmation, but we are presently discussing a mind somewhat more primitive than that.

A human is veritably supersaturated with cognitive forces that would let us make deductions about Travelness, and any mind-in-general will be saturated with information about how a low-entropy universe works. However, it is still possible that the generalizations will stop completely short of Travelness—that the AI will not only regard hypotheses about Travelness coherence as *low probability,* but as *entirely unsupported by any known piece of sensory information,* because Travelness is believed to be made out of different stuff.

Is there *any* imaginable configuration of an AI's mind that will naturally avoid all hypotheses about Travelness? I think so, but it's a pretty exotic configuration. The AI has to believe—in the sense of seeing no reason to believe otherwise—that Travelness is temporally noncoherent, that Travelness has noncoherent content, and that the programmers' use of a general statement to produce specific statements is irrelevant to the content of Travelness. Travelness being defined as "What programmers say," if translated into a specifically exotic model, could meet these requirements. The AI would need to assume that the Travelness referent changes at the exact moment when the program-

mer's lips move, and that any cognitive or physical processes that take place previous to the programmer's lips moving are irrelevant (insofar as sensory information about Travelness is concerned). Under those circumstances, all of the above methods for reasoning about Travelness would, in fact, be incorrect; would produce incorrect predictions. Of course this definition can very easily short-circuit, as depicted in 5.5 Interlude: Why Structure Matters.

In fact, it looks to me like any definition which does not enable probabilistic reasoning about the supergoal referent must define the referent as incoherent both temporally and "spatially," and must define the referent as being identical with the sensory information produced about it (or rather, becoming identical at the instant of production of such information, then remaining identical thereafter).

I myself evaluate a very high probability that an AI would *somehow* wind up with expectations about Travelness unless the programmers made a deliberate attempt to prevent it. For example, even given an exotic structure which permits no expectations about Travelness in advance of sensory information, the AI could still evaluate a finite (albeit very, very small) probability that sensory information had been produced but dropped, or erased from memory by later intervention, and so on. In the absence of diabolic misinterpretation, of course, this is a nearly infinitesimal probability and will generate nearly infinitesimal desirability differentials, but still. Technically, the AI is trying to correct mistakes in the transmission of sensory data, rather than forming expectations about supergoal content, so this doesn't really count from a Friendly-AI structural standpoint. It does, however, show how hard it is to develop *totally* incoherent supergoals. Similarly, even an AI with a "short-circuited" definition of "Travelness" might conclude that the programmer's lips are likely to move and thereby alter Travelness in a certain way, and move vehicle 4145 into position in anticipation of greater future supergoal fulfillment; this is sort of half-way between the two possibilities as far as relevance is concerned.

The atomic definition of convergence, as you'll recall from 3.1 Interlude: The Story of a Blob—what, you say? You don't recall what was in that topic? You're not even sure that it was in the same paper? You can't recall anything from before you started reading 5 Design of Friendship Systems, including your own childhood? I guess you'll just have to start reading again after you've finished. Anyway, the atomic definition of convergence is when a system makes the same choice given two different conditions; a blob turning to swim *towards* nutrients, regardless of whether the blob was originally swimming east or west; a "mathblob" adding 2 to 65 and 3 to 64, to achieve 67 in either case.

If a Travel AI with external reference semantics has any expectations at all that "Travelness" is a *something*, a thing in a low-entropy Universe that at least *might* obey some of the same rules as other things, then the Travel AI will form expectations about Travelness. This doesn't require that Travelness be defined as a sentience-independent physical

object floating out in space somewhere; all that's required is that Travelness have some definition that's physically derived or that has some connection to our low-entropy Universe. If you show the Travel AI four thousand similar statements, they'll have at least a *little* inertia, a *little* effect on the four-thousand-and-first.

This small degree of convergence doesn't prove that the Travel AI will suddenly break free of all human connections—if the Travel AI has a 1%-confidence belief in a 1%-strength correlation, the differential desirabilities are pretty small compared to higher-priority content. Even so, it would be possible for a programmer to deliberately "damp down" the asserted confidence of some piece of programmer-derived sensory information—tell the AI, "With a confidence of 0.001%, vehicle 4145 should move at 39 mph"—and, in the absence of injunctions, the Travel AI will still think it more likely that vehicle 4145 should move at 40 mph. That is, the Travel AI will think it most likely that vehicle 4145 should move at 40 mph, regardless of whether (binary branch) the programmer says "With a confidence of 0.001%, vehicle 4145 should move at 40 mph," or "With a confidence of 0.001%, vehicle 4145 should move at 39 mph." This is a tiny, tiny amount of convergence, and a tiny, tiny amount of programmer independence[71]—but it's there.

5.6.1.2 Some Forces That Shape Friendliness: Moral Symmetry, Semantics of Objectivity

The basic cognitive structure for external reference semantics leaves unspecified where Friendliness content actually comes from; so, as depicted above, it's possible to come up with different definitions, and different evaluated probabilities based on the different definitions. A "Friendly AI" with external reference semantics *and nothing else* may not contain any information at all that would help the Friendly AI make a decision about *where* the programmers get their knowledge about supergoals. A positive outcome would be if the Friendly AI assumed that the programmers, who know about Friendliness content, are also the most reliable source for information about where Friendliness comes from, and thus accepted the programmers' statements as sensory information. However, without *a priori* knowledge or causal validity semantics, this generalization would have to be made blindly.[72]

71. Causally speaking, of course, the AI's decision has still been completely determined from programmer-provided data; the "programmer independence" here is an infinitesimal amount of independence from *additional* programmer interventions.

72. In theory, an AI built solely with external reference semantics might accept the programmers' statements as the only available information and therefore create within itself causal validity semantics—but accept the programmers' statements at only 10% probability. In which case you might be stuck, from now

As with Friendliness supergoal content itself, the issue of "Where does Friendliness come from?" is complex enough that no snap answer should be embedded as "correct by definition" in the AI. We can thus immediately see that the reply to "Where does Friendliness come from?" requires a method for learning the answer, rather than a snap answer embedded in the code and enshrined as correct by definition. Similarly, external reference semantics provide a method for growing and correcting whatever interim answer is being used as Friendliness content. However, this section, "shaper/anchor semantics," is just about the interim answer used by *Creating Friendly AI* for "Where does Friendliness come from?", in the same sense that *Creating Friendly AI* uses the volition-based definition when talking about Friendliness content. It's the later section 5.6.2 Causal Validity Semantics that closes the loop, explaining how, e.g., a Friendly AI with only external reference semantics could acquire shaper/anchor semantics.

As humans, of course, we leave the factory with built-in causal validity semantics. These intuitions will now be applied to the question at hand: Where does Friendliness come from?

Supposing I were to ask you a question about Friendliness, where would you get your answer? What are some of the forces that might affect your answer? Suppose, for example, that someone were to propose to you that the request of a human whose last name ends in a 'p' is worth only half as much (has half as much desirability) as the request of a human whose last name ends with any other letter.[73] In this case, the primary reason for your instant rejection is fairly simple: You don't perceive any cause whatsoever for that modification, and you have a minimum-encoding, low-entropy perception of Friendliness.

Suppose the suggester presses her case—for example, by offering a reason such as "Well, I'm human, a volitional entity, and what I want matters, and I want a Friendly AI to dislike people whose last names start with 'p'—in fact, I'll experience mental anguish if I live in a world where Pfolk are equal citizens." In this case, you would probably bring

---

until the end of time, with a "Friendly AI" that behaves exactly as if ve had assigned your statements 90% or 99% probability, since there are no other sources of information, but that *actually* only assigns your statements 10% or even 1% probability. Since the Friendly AI would be effectively emulating one with real causal validity semantics, the "tail" end, the original external reference AI that accepted the information, might be protected by uncertainty from modification and solidification by the emulated causal validity semantics. The uncertainty might even be high enough to break 5.3.3 Unity of Will.

Today's dose of humor edged with adrenaline: Imagine a *Sysop* that only thinks you have a 1% chance of being right, but has absolutely nothing else to do . . .

Moral of the story: To prevent a philosophical crisis, implement causal validity semantics *before* the AI can self-modify, or *before* the AI improves past unity of will.

73. Pfolk discrimination applies only to those whose last names use Latin lettering, of course.

up an argument to the effect that all people (all humans, all sentient beings, et cetera) should be treated as morally equal, which you regret to announce is an overriding factor as far as you're concerned. (I would say the same thing, by the way.) If the suggester presses the case further, it will probably be by announcing that the Pfolk were shown to be responsible for 90% of all mime performances, and therefore deserve whatever they get. Let's stop the argument here for a moment, and try to look at some of the underlying forces.

Moral equality is not only a powerful ambient meme of the post-WWII era, but also a very direct manifestation of a panhuman cognitive pressure towards moral *symmetry.* This is probably not the best term, since it's very close to "moral equality," but it's the best one I can offer. Moral symmetry is supported by three cognitive forces; first, the way we model causality; second, our having evolved to persuade others; third, our having evolved to resist persuasion.

For humans engaged in moral argument, everything needs to be justified. We expect a cause for all decisions as we expect a cause for all physical events. Someone who doesn't think that a decision requires a cause is not only cognitively unrealistic—it's really hard to imagine acausal anything—but subject to exploitation by anyone with a more coherent mind. "Could you give me all of your money?" "Why should I?" "There's no reason whatsoever why you should." "Okay, I'll do it!" Similarly, someone who believes that decisions are acausal is likely to be an ineffective persuader: "Could you give me all of your money?" "Why should I?" "There's no reason why you should." "Ummm . . . no."

In the discussion on "moral deixis," the example was given of John Doe saying to Sally Smith, "My philosophy is: Look out for John Doe," and Sally Smith hearing, "Your philosophy should be: Look out for Sally Smith," rather than hearing: "Your philosophy should be: Look out for John Doe." The conclusion there was that we have very strong expectations of speaker deixis and automatically substitute the [speaker] variable for any heard self-mention. The conclusion here is that if John Doe expects Sally, like himself, to have a built-in instinct for the protection of John Doe, John is doomed to disappointment. For John Doe to be an effective persuader, he must make use only of cognitive forces that he has a reasonable expectation will exist in Sally's mind. He can send an argument across the gap either by recasting his arguments to appeal to Sally's own observer-centered goals, or by using the semantics of objectivity.

The latter really ticks off the moral relativists, of course. Moral relativists insist that no objective standard of morality exists, and that arguments that use the semantics of objectivity are automatically flawed, thereby appealing to the universal human preference for unflawed arguments; they then go on to use moral relativism to argue against some specific moral principle as being ultimately arbitrary, thereby appealing to the universal human prejudice against arbitrariness. Um, full disclosure: I hate moral relativism with

a fiery vengeance, and I hate cultural relativism even more, but rather than going on a full-scale rant, I'll (for the moment) just state my position that any public argument is, de facto, phrased in terms which appeal to a majority of listeners. If a moral relativist wants to appeal to an audience prejudice against arbitrariness by saying that all morality is arbitrary and therefore Friendliness is arbitrary, I'm justified in using the criteria of that audience prejudice against arbitrariness as my objective standard for arguing whether or not Friendliness is arbitrary.

This doesn't prove that total moral relativism is logically inconsistent; (honest) evangelism of total moral relativism is logically inconsistent, but it's theoretically possible that there could be millions of logically consistent, honest, total moral relativists keeping their opinions private. However, *arguing* with me in front of an audience about moral relativism only makes sense relative to some agreed-upon base layer held in common by the audience and both debaters; all I need to do is show that Friendliness meets the criterion of that base layer. See 5.6.2.6 Objective Morality, Moral Relativism, and Renormalization below.

Anyway, the upshot is that, for any sort of diverse audience, humans generally use the semantics of objectivity, by which I mean that a statement is argued to be "true" or "false" without reference to data that the audience/persuadee would cognitively process as "individual." (Whether the appealed-to criterions are human-variant data that the audience happens to have in common, or panhuman complex functional adaptations, or characteristics of minds in general, or even a genuine, external objective morality, is irrelevant to this particular structural distinction.) Appeals to individualized goals are usually saved for arguing over what kind of pizza to get, or convincing someone to be your ally in office politics, and so on—individual interactions, or interactions with a united audience. Thus, when humans talk about "morality," we generally refer to the body of cognitive material that uses the semantics of objectivity.

This holds especially true of any civilization that's been around long enough to codify the semantics of objectivity into a set of declarative philosophical principles, or to evolve philosophical memes stating that observer-centered goals are morally wrong. Even if some subgroup within that civilization (Satanists, moral relativists, Ayn Rand's folk) has a philosophy that makes explicit reference to observer-centered goals, the philosophy will have an attached justification stating, in the semantics of objectivity, the reason why it's okay to appeal to observer-centered goals. Anyone who grows up in a civilization like that is likely to have a personal philosophy built from building blocks and structure that grounds almost exclusively in statements phrased in the semantics of objectivity, and has a reasonable expectation that a randomly selected other citizen will have a similarly constructed philosophy, enabling the semantics of objectivity to be used in individual interactions as well.

The semantics of objectivity are also ubiquitous because they fit very well into the way our brain processes statements; statements about morality (containing the word "should") are not evaluated by some separate, isolated subsystem, but by the same stream of consciousness that does everything else in the mind. Thus, for example, we cognitively expect the same kind of coherence and sensibility from morality as we expect from any other fact in our Universe.

In the example given at the start of this subsection, someone had just proposed discrimination against the Pfolk; that the request of a person whose last name starts with "p" should be valued (by a Friendly AI) at one-half the value of any other citizen. So far, the conversation has gone like this: "Discriminate against the Pfolk." "Not without a reason." "I'll be unhappy if you don't." "That's not a reason strong enough to override my belief in moral equality." "Pfolk are responsible for 90% of all mime performances, so they deserve what they get."

In that last parry, in particular, we see an appeal to moral symmetry. Moral symmetry is a cognitive force, not a moral principle (the moral principle is "moral equality"), but if we were to try and describe it, it would go something like this: "To apply an exceptional moral characteristic to some individual, the exceptional moral characteristic needs to be the consequence of an exceptional attribute of the individual. The relation between individual attribute and moral characteristic is subject to objectivity semantics."[74] There's a very strong cognitive pressure to justify philosophies using justifications, and justifications of justifications, that keep digging until morally symmetric, semantics-of-objectivity territory is reached.

There's an obvious factual component to the statement "Pfolk are responsible for 90% of all mime performances, so they deserve what they get"—as far as I know, Pfolk are *not* responsible for 90% of all mime performances, not that I've checked. In this case, the factual reference is very near the surface; however, factual references quite often pop up, not just during moral debates, but during philosophical debates (in discussions about how to choose between moralities). This, again, is a consequence of our brains using the same semantics of objectivity for facts and morality; the very distinction between "facts" and "morality" (or "supergoals" and "subgoals," for that matter) is a highly sophisticated

74. That last part, if not directly asserted to be a universal rule, is subject to recursion until it grounds in a tail-end moral symmetry: If the relation between attribute and characteristic is unique to that individual (rather than general to everything cognitively processed as a person), the uniqueness is also justified using objectivity semantics, i.e., either the exception is justified by reference to some attribute, or a blanket statement is made that arbitrary exceptions are permitted. That's only if the declarative justifications reach that far down, of course; but if not, and you ask the philosopher to justify herself, she will invent—on the fly—one of the justifications listed.

discrimination, so it's not surprising that the two are mixed up in ordinary discourse. This is not important to the present discussion, but it will become important shortly.

We've now seen several factors affecting our beliefs about Friendliness, our beliefs about supergoals, and our beliefs about morality (communicable supergoals). Some of them are high-level moral beliefs, such as moral equality. Some of them are more intuitive, such as moral symmetry and our tendency to "put yourself in the other's shoes." Some lie very close to the bottom layer of cognition, such as our using a single brainwide set of causal semantics for all thoughts, including thoughts about morality.

5.6.1.3 Beyond Rationalization

We use the whole of our existing morality to make judgements about the parts. Of course, since we're humans, with observer-biased beliefs and so on, this trick often doesn't work too well. However, so long as you have enough seed morality to deprecate the use of observer-biased beliefs, and you happen to be a seed AI with access to your own source code, "nepotistic" self-judgements should not occur—that is, if the system-as-a-whole has a valid reason to make a negative judgement of some particular moral factor, then that negative judgement (modification of beliefs) will not be impeded by the current beliefs. Nor will the fact that some particular judgement winds up contradicting a previously "cherished" (high confidence, high strength, whatever) be experienced as a cognitive pressure to regard that judgement as invalid—that's also a strictly human experience.

A default human philosophy (i.e., one that operates under the evolved design conditions) is a system that interestingly contradicts itself. (Note that the word is *default,* not *average.*) A default human will phrase all his moral beliefs using the semantics of objectivity (for the reasons already discussed), but trash all actual objectivity through the use of observer-biased beliefs—a complex process of "rationalization," whereby conclusions give birth to justifications rather than the other way around. Because of this rationalization-based disjunction between morality and actions, the two have been, to some degree, pushed around *independently* by evolution. As long as rationalization (at sufficient strength and reliability) is already present as an adaptation, evolution can freely modify moral reasoning to use the semantics of objectivity, or justification by public benefit, without causing the organism to actually make objective decisions or act for the public benefit, both of which might be a disadvantage (from evolution's perspective, anyway). (As always in discussions of evolutionary psychology, readers should bear in mind that evolutionary motives are not the same as cognitive motives, and that evolutionary hypocrisy does not imply cognitive hypocrisy. All actual *actions* being equal, cognitive hypocrisy is a disadvantage in social situations; humans are imperfectly deceptive.)

Actually, rationalization does not *totally* disjoint morality and actions; it simply gives evolution a *greater* degree of freedom by *loosely* decoupling the two. Every now and then, the gene pool or the memetic environment spits out a genuine altruist; who, from evolution's perspective, may turn out to be a lost cause. The really *interesting* point is that evolution is free to load us with beliefs and adaptations which, if executed in the *absence* of rationalization, would turn us into total altruists ninety-nine point nine percent of the time. Thus, even though our "carnal" *desires* are almost entirely observer-centered, and our social desires are about evenly split between the personal and the altruistic, the adaptations that control our *moral justifications* have strong biases toward moral symmetry, fairness, truth, altruism, working for the public benefit, and so on.

Evolution, however, being blind, foresaw neither the printing press nor the Internet. Rather than social morality and personal rationalization being equally balanced, those philosophies that are socially acceptable now have a chance to build up over time. Since the second half of the twentieth century, in particular, humanity has accumulated a lot of ambient memes explaining how to overcome rationalization. The conflict between rationalization-corrupted altruism and genuine altruism is now proverbial; equally, our conscious tendency to believe that we are altruistic automatically biases us towards thinking that a way to overcome rationalization is a good thing. Humanity as a whole has not yet *won* the conflict between moral intuitions and rationalization, but the balance is definitely not what it was in the ancestral environment.

A human trained in evolutionary psychology can pretty much win completely, as long as he starts his journey with the surface belief that altruism is good and rationalization is bad. For an AI, there isn't even a contest—again, as long as the *seed philosophy* starts out with a surface belief validating altruism and deprecating rationalization. It shouldn't even matter where this belief comes from, so long as it's there.

5.6.1.4 Shapers of Philosophies

shaper. A shaper is a philosophical affector, a source of supergoal content or a modifier for other shapers; a belief in the AI's philosophy; a node in the causal network that produces supergoal content.

The essential idea underlying shaper/anchor semantics is that, rather than transferring the *outputs* of our philosophies to the AI as supergoal content, we also transfer the *philosophies themselves,* so that the AI can guess our responses, produce new supergoal content and revise our mistakes. This doesn't just mean the first-order causes of our decisions, such as moral equality, but the second-order and third-order causes of our decisions, such as moral symmetry and causal semantics. Shapers can validate or deprecate other shapers: For example, memetic survival rates play a large part in morality, but we're likely to think that the differential survival of memes using objectivity seman-

tics is a "valid" shaper, while deprecating and trying to compensate for the differential survival of memes appealing to hatred and fear. Rationalization is strongly deprecated; false factual beliefs acting as shapers are even more strongly deprecated.

Anchors are described in more detail below and allow the AI to *acquire* shapers by observing humans, or by inquiring into the causes of current philosophical content.

Shaper/anchor semantics serve the following design purposes:

- They eliminate (or make reversible) a major class of programmer errors, allowing convergence to "truly perfect Friendly AI" over an entire range of mistakes.
- They vastly enhance programmer independence.
- They ground the first-order external reference semantics.

SAS: Correction of Programmer Errors

A generic goal system can detect and correct programmer errors in reasoning from supergoals to subgoals.

A shaper-based goal system can detect and correct programmer errors where the error is made explicit—for example, where the programmer says "Moral principle B is the result of shaper A," and it's not—the programmer is (detectably) engaging in cognitive activity of a type labeled by the current system as "rationalization," or is making a conclusion influenced by beliefs that are factually incorrect.

An anchor/shaper Friendly AI can detect and correct implicit errors—for example, where the programmer says "A is moral," the AI deduces that the statement is made as a result of within-the-programmer shaper B, plus a factual error C. The Friendly AI can assimilate B, correct the factual error C (producing correct factual belief D), and use B plus D to produce the *correct* moral statement E. (There are several emendations to this general principle, both under anchoring semantics and under causal validity semantics; see below.)

SAS: Programmer-Independence

Thanks to the way human society works, humans have a strong need to justify themselves. Thanks to the memetic rules that govern an interaction between any two people, each time you ask "Why?", the farther down you dig—the more likely the person is to attempt to justify herself in terms that are universal relative to the target audience. Sometimes these universals are cultural, but, since nobody makes a *deliberate* effort to use *only* cultural justifications, sometimes these universals happen to be panhuman attributes, or even very deep attributes like causal semantics (which could plausibly be a property of minds in general). Very often the justifications are rationalizations, of course, but that

doesn't matter; what matters is that as long as the AI learns the *given* justifications instead of the surface decisions, the AI will tend to wind up with a morality that grounds in human universals—certainly, a philosophy that's a lot closer to grounding in human universals than the philosophy of whichever human was used as a source of philosophical data. Another way of phrasing this is that if you seat two different humans in front of two different AIs with complete structural semantics, you'll tend to wind up with two AI philosophies that are a *lot* closer than the philosophies of the two humans.

Also, of course, the surface of the philosophy initially embedded in a competently created Friendly AI would have a strong *explicit* preference that validates programmer-independence and deprecates programmer-dependence. An extension of this principle is what would enable a Friendly AI to move from exhibiting normative human altruism to exhibiting the normative altruism of *minds in general,* if the philosophy at any point identifies a specific difference and sees it as valid.

> Note: When I say the "surface" of the philosophy, I refer to the proximate causes that would be used to make any given immediate decision. It doesn't imply *shallowness* so much as it implies *touching the surface,* if you think of the philosophy as a system in which causes give rise to other causes and eventually effects; a really deep, bottom-layer shaper can still be "surface"—can still produce direct effects as well as indirect effects.

If you seat two different humans with an explicit, surface-level preference for programmer-independent Friendliness in front of two AIs with complete structural semantics, you will quite probably wind up with two identical AIs. (In later sections I'll make all the necessary conditions explicit, define programmer-independence, and so on, but we're getting there.)

SAS: Grounding for External Reference Semantics

In a later section, I give the actual, highly secret, no-peeking target definition of Friendliness that is sufficiently convergent, totally programmer-independent, and so on. Hopefully, you've seen enough already to accept, as a working hypothesis, the idea that a philosophy can be grounded in panhuman affectors. The programmers try to produce a philosophy that's an approximation to that one. Then, they pass it on to the Friendly AI. The Friendly AI's external referent is supposed to refer to that programmer-independent philosophy, about which the programmers are good sources of information, as long as the programmers give it their honest best shot. This is *not* a complete grounding - that takes causal validity semantics—but it does work to describe all the ways that external reference semantics should behave. For example, morality does *not* change when words leave the programmers' lips, it is possible for a programmer to say the wrong thing,

the cognitive cause of a statement almost always has priority over the statement itself, manipulating the programmer's brain doesn't change morality, and so on.

(Note also that an AI with shaper semantics cannot nonconsensually change the programmer's brain in order to satisfy a shaper. Shapers are not meta-supergoals, but rather the causes of the current supergoal content. Supergoals satisfy shapers, and reality satisfies supergoals; manipulating reality to satisfy shapers is a non-sequitur. Thus, manipulating the universe to be "morally symmetric," or whatever, is a non-sequitur in the first place, and violates the volition-based Friendliness that is the *output* of moral symmetry in the second place.)

(Of course, if the anticipation of supergoal satisfaction is set up in the wrong way, and the definition of Friendliness is at least partially grounded in the human species, it is conceivable that an ends-justify-the-means case of manipulating the human species' beliefs about morality, in anticipation of greater "supergoal satisfaction" later on, could short-circuit the goal system. More about that later. But at least it would *be* a case of ends justifying means; under meta-supergoals, where the supergoal of avoiding nonconsensual mindspace violation is a strict subgoal of maximizing human satisfaction with the supergoals, the short-circuit isn't even seen as *temporarily* undesirable.)

5.6.1.5 Anchors

anchor. Anchor semantics are a structural attribute that enable the AI to discover and absorb philosophical content even if the programmers themselves are unaware of it. An "anchor" is an element of the AI's current philosophical content—supergoal content, subgoal content, or shaper content—whose causal antecedents are not fully known.

An anchor happens when the human says "I think X is a good idea" but isn't clear about *why,* or isn't sure she is fully aware of all the valid factors that played a part in the decision. An anchor is a reference to whichever cognitive events were occurring at the space-time event that was the human making the decision to add the anchor. Dereferencing the anchor will probably be carried out by (externally) asking the programmer probing questions, or by (internally) using Bayesian reasoning, Occam's Razor, known facts about humans, et cetera. (Of course, internal reasoning of this type may tend strongly towards the probabilistic rather than the near-certain.) In theory, a Transition Guide—a *successful* seed AI turned superintelligence—could initiate a nondestructive brain scan of the programmer's neural mind state and find out *all* the cognitive causes of the human's belief in the anchoring point. I mention this possibility mainly to show that an anchor is a reference to information that is *entirely* objective—a physical fact. However, the *referent* of the anchor, once the anchor is successfully dereferenced, be-

comes standard information about humans—i.e., it has as much (or as little) effect as would a direct statement by the human.

Positive Anchors

A positive anchor is an action or state marked as anchor-desirable; a positive anchor is the AI's conclusion about probabilistic shaper/supergoal/subgoal content, made from the sensory information that is a human stating "This action/state is desirable [for reasons X], but I'm not completely sure why [that those are all the reasons]."

A positive anchor would enable (for example) the humans to ask the AI to do something that is clearly a subgoal rather than an end in itself, but which was not a subgoal of any currently known (to the AI) supergoal content, nor a subgoal of the unknown supergoal content that supports subgoals regarding the goal system itself. This could also apply to shaper content or supergoal content. If the humans felt that something was a good idea, but not an end in itself, and weren't sure why, it would become a positive anchor. I strongly feel that a positive anchor should be used with caution to ensure that it behaves as a reference to the causes that *were in fact in the minds of the humans,* and not as a rationalization target that promotes *any* line of reasoning that happens to intersect it.

A positive anchor should be distinguished from an ordinary programmer affirmation. An anchor is when *the humans don't know,* or aren't sure they know completely. A programmer affirmation is when the AI is too young to understand the justification, or the humans just haven't got around to explaining it. A positive anchor is a request that the AI try to discover the philosophical causes on vis own. A programmer affirmation is a request to trust the programmers today until the explanation is provided tomorrow. Under vanilla external reference semantics, without any information about shapers, *all* supergoal content is programmer-affirmed. (However, an AI that has shaper/anchor *semantics* but is provided only with external-reference *content* still has the philosophical capability to hypothesize a line of reasoning under which the programmers are "wrong," just as a general intelligence acting entirely on programmer-affirmed data would possess the same capability with respect to hypotheses about the physical world.) ERS content might also be analogized to a programmer-affirmed subgoal supported entirely on trust (i.e., rather than supported by a programmer-affirmed link to a parent goal)—the programmers saying "We think this is a good idea, and you know we more than you do, and you don't have any reason to believe we're wrong, so go ahead and do it—it's more likely to work out than not."

Negative Anchors

A negative anchor is an action or state marked as anchor-undesirable; a negative anchor is the AI's conclusion about probabilistic shaper/supergoal/subgoal content, made from the sensory information that is a human stating "This action/state is undesirable [for reasons X], but I'm not completely sure why [that those are all the reasons]."

Like a positive anchor, the power of a negative anchor is to lend greater resilience to a chain of logic. Suppose there's a negative subgoal marked as a negative anchor. Even if a key justification—dimly understood by the programmers, but not represented at all by the AI—is left out, and the humans originally justified the negative subgoal using a *fallacious argument,* the collapse of that fallacious argument will not completely extinguish the negative anchor. Instead, the AI will hypothesize, at whatever degree of probability, that there is an unknown but valid reason for the negative anchor, and go looking for it.

Rationalization *away from* negative anchors seems less inherently dangerous than rationalization *towards* positive anchors, but psychological realism is still an important constraint.

Note also that a Friendly AI is capable of *hypothesizing* that something has, or should have, anchor status—i.e., that another shaper exists behind it, or that it's a good idea to look for such a shaper—even if the humans don't tell ver so.

Anchor Abuse

Programmer affirmation of anchor status, like any other programmer affirmation, is dependent on the AI's belief that the programmers are likely to be correct—in Bayes logic, that the programmer has a small chance of offering an anchor in a case where no anchor is necessary.

Anchors, like any assertions, have probabilities and strengths. Because of our human absolutist psychology, it's tempting to take something that we *really don't want the AI to discard* and assign it maximum-probability, maximum-strength anchor status—for example, to make "Don't kill humans" a maximum-probability negative anchor. One, this is adversarial. Two, this is confusing negative anchors with ethical injunctions. Three, this is *lying to the AI*—technically, making a mistake that has a high probability of being a common cause for lots of other mistakes—and goes that little extra length towards precipitating a philosophical crisis.

Negative anchors, as a tool, exist when there is *uncertainty on your part* that you know all the reasons why something is undesirable, or when you have cause to believe there's a reason that isn't on the list. If you think that there would be a *large negative penalty* for

forgetting a reason, causing the AI to incorrectly deprecate [whatever], then that's an *injunction* which says that [whatever] should be treated as having negative anchor status.

Ultimately, what you say is simply sensory information to the AI. If your nervousness about some negative point being violated causes you to fill the air with negative anchors, then the AI will, quite correctly, deduce that all the negative anchors are a result of your nervousness. So if the AI has some reason (technically, the Bayesian prior, independent of programmer affirmations) to think that your nervousness is wrong, then that hypothesis would invalidate all the anchors as well—or rather, would invalidate all the sensory information produced as a consequence of your nervousness, no matter how strong or how high a probability you asked the AI to assign to the negative anchors. If you honestly think logic of this type is more likely to indicate failure of Friendliness on the AI's part than failure of normative altruism on your own part, then that is an *ethical injunction,* not a negative anchor.

5.6.1.6 Useful Shaper Content Requires High Intelligence

Shapers take place in a domain—philosophy—that is one of the most sophisticated domains around. It is possible that even a young AI will be able to manipulate information about shapers in such a way to make useful deductions in advance of programmer-provided information, and even possible that a young AI will be able to correctly second-guess the programmers. But I wouldn't bet on it. I would be less surprised to find the young AI's predictions being totally useless, or right only by coincidence.

It might be wise to try and teach an AI shaper content before vis intelligence has reached the necessary level, as long as everything is running at very low intrinsic probabilities, so that the AI simply gains experience by being proved wrong over and over again—still useful data. However, this does mean that it may be harder to *test* shaper content (philosophy) at this stage—to verify that it actually *does* work, as opposing to doing as much untested work in advance as possible. Depending on the intrinsic structure of the hard takeoff trajectory, this may mean that it's impossible to achieve supersaturated Friendliness in this area in advance of the point in time where a hard takeoff becomes possible. In fact, it may even mean that shaper content is still basically useless—insufficient intelligence to be reliable at all, much less human-competitive—at the point where a hard takeoff first becomes possible. I'll talk about how to solve this class of problem later (in 5.7 Developmental Friendliness).

However, *abstract* reasoning about shaper *semantics,* without reference to *specific* shapers, may still be enough to establish important early beliefs. For example, the idea that a later programmer knows more than the same programmer at an earlier point in time, and hence that later statements should override earlier ones, requires understanding that the programmer is not a black box; that the programmer can make a "mistake"

and "correct" it. Thus, an AI may have need of shaper *semantics* in an early stage of development, even if shaper *content* must wait on the acquisition of sufficient intelligence.

5.6.2. Causal Validity Semantics

Causal validity semantics subsume both external reference semantics and shaper/anchor semantics. Causal validity semantics:

- Define the ultimate source of validity for the goal system—the "reason why the AI cares."
- Describe a specific method of anchoring the AI's entire goal system in data external to the AI (not external to humanity, perhaps, but external to the AI).
- Provide a framework for generalized architectural changes to the goal system as well as alteration of supergoal content.
- Provide a framework for defining and achieving complete programmer-independence.

5.6.2.1 Taking the Physicalist Perspective on Friendly AI

An AI's complete mind-state at any moment in time is the result of a long causal chain. We have, for this moment, stopped speaking in the language of *desirable* and *undesirable,* or even *true* and *false,* and are now speaking strictly about *cause* and *effect.* Sometimes the causes described may be beliefs existing in cognitive entities, but we are not obliged to treat these beliefs *as beliefs,* or consider their truth or falsity; it suffices to treat them as purely physical events with purely physical consequences.

This is the physicalist perspective, and it's a dangerous place for humans to be. I don't advise that you stay too long. The way the human mind is set up to think about morality, just imagining the existence of a physicalist perspective can have negative emotional effects. I do hope that you'll hold off on drawing any philosophical conclusions until the end of this topic at the very least.

That said . . .

The complete causal explanation for any given object, any given spacetime event, is the past light cone of that event—the set of all spacetime events that can be connected, by a ray of light, to the present moment. Your light cone includes the entire Earth as of one-eighth of a second ago, any events that happened on the Sun eight-and-change minutes ago, and includes the Centauri system after four-and-change years ago. Your current past light cone does not include events happening on the Sun "right now,"[75]

75. That is, your present space of simultaneity.

and will not include those events for another eight-and-change minutes[76]; until a ray of light can reach you from the Sun, all events occurring there are causally external to this present instant.

The past light cone of a Friendly AI starts with the Big Bang. Stars coalesce, including our own Sol. Sol winds up with planets. One of the planets develops complex organic chemicals. A self-replicating chemical arises. Evolution begins, as detailed (a bit metaphorically) in 3.1 Interlude: The Story of a Blob. The first convergent behaviors arise—except that under the physicalist perspective, they are neither "convergent," nor "behaviors." They happen a certain way, or they don't. They are simply historical facts about blobs and genes. If one were to go so far as to abstract a description from them, it would consist of statistical facts about how many events fit certain descriptions, such as "blob swims towards nutrients."

The blobs grow more complicated, nervous systems arise, and goal-oriented behaviors begin to give way to goal-oriented cognition. Entities arise that represent the Universe, model the Universe, and try to manipulate the Universe towards some states, and away from others. Eventually, sentient entities arise on a planet. In at least one sentient, the representation of the Universe suggests that it can be manipulated toward certain states and away from other states by building something called "Friendly AI" and imbuing this AI with a model of the Universe and differential desirabilities. The sentient(s) carry out the actions suggested by this belief and a Friendly AI is the result.

The Friendly AI has a final cognitive mind-state (that is, a final set of physically stored information) which is causally derived from the Friendly AI's initial mind-state, which is causally derived from the programmers' keystrokes—

> Note: A word about terminology: When using the physicalist perspective, it's important to distinguish between *historical dependency* and *sensitivity to initial conditions.* If two different Friendly AIs have different initial states and converge to the same outcome, they can have a completely different set of historical *dependencies* without having any subjunctive *sensitivity.* To put it another way, when using the physicalist perspective, we are concerned simply with who *did* hit the keystrokes in the causal chain. A Friendship programmer seeking "convergence" cares about whether a different person "would have" hit the same keystrokes. But the physicalist perspective can only describe what *actually happened.*

76. Assuming that none of my readers are traveling at a significant fraction of lightspeed—i.e., that their duration lengths and spaces of simultaneity approximate my own.

—which are causally derived from the programmers' mental model of the AI's intended design. The causal source of goal system architecture (again, without reference to *sensitivity*) is likely to be, almost exclusively, the programming team. The ability of that programming team to create the first AI is likely to have been, historically, contingent on choices made by others about funding and support.

The physicalist perspective observes *all* events, and even the *absence* of events, within the past light cone. So the ability of the programming team to create the first AI can also be said to be dependent on any external individuals who choose *not* to interfere with the AI, dependent on the fact that the project's supporters had resources available, and so on. (The *existence* of an Earthbound AI would be contingent on other events and nonevents as well, such as an asteroid not crashing into the Earth and the fact that the Sun has planets, but the *specific* content of the AI does not seem to be dependent on such events.)

The specific differential desirabilities and goal system architecture given/suggested to the Friendly AI—again, with reference only to *history* and not *sensitivity*—are causally derived from the surface-level decisions of the programming team in general and the Friendship programmers in particular. A given surface-level decision is produced by the programmer's observe-model-manipulate cognitive process. A series of keystrokes (lines of code, programmer affirmations, whatever) is formulated which is expected to fulfill a decision or subgoal, after which a series of motor actions result in a series of keyboard inputs being conveyed to the AI.

The surface-level decisions of the programming team are the causal result of those programmers' mind-states, the primary relevant parts of which are their "personal philosophies." The *complete* mind-states include panhuman emotional and cognitive architecture, and bell-curve-produced ability levels[77]; iteratively applied to material absorbed from their personal memetic environments, in the process of growing from infant to child and child to adult.

I will now analyze the sample case of the development of my own personal philosophy in more detail. Just kidding.

77. As amended for technological neurohacking, BCI, the use of neuroimaging Mirrors (if that technology matures), and any Friendship programmers who happened to be born with unusual neurologies that made them well-suited to AI research. Genetic engineering would also be included if the Singularity waits long enough (a) for the technology to advance to the point where genetically enhanced intelligence is discernibly different from just having smart parents, and (b) for the kids to grow up; this probability is effectively zero.

5.6.2.2 Causal Rewrites and Extraneous Causes

We are temporarily done with the physicalist perspective. Back to the world of desirable/undesirable, true/false, and the perspective of the created AI.

The basic idea behind the human intuition of a "causal validity" becomes clear when we consider the need to plan in an uncertain world. When an AI creates a plan, ve starts with a mental image of the intended results and designs a series of actions which, applied to the world, should yield a chain of causality that ends in the desired result. If an *extraneous cause* comes along and disrupts the chain of causality, the AI must take further actions to preserve the original pattern; the pattern that would have resulted if not for the extraneous cause.

Suppose that the AI wants to type the word "friendly." The AI plans to type "f," then "r," then "i," et cetera. The AI begins to carry out the plan; ve types "f," then "r," and then an extraneous, unplanned-for cause comes along and deposits a "q." Although it may be worth checking to make sure that the extraneous letter deposited is not the desired letter "i," or that the word "frqendly" isn't even better than "friendly," the usual rule—where no specific, concrete reason exists to believe that "frqendly" is somehow better—is to *eliminate the extraneous cause to preserve the valid pattern.* In this case, to alter the plan: Hit "delete," then hit "i," then "e," et cetera.

Cognitive errors are also extraneous causes, and this applies to both the programmer and the AI. If the programmer types a "q" where "i" is meant, or writes a buggy line of code, then an extraneous cause has struck on the way from the programmer's intentions to the AI's code, or system architecture. If a programmer designs a bad Friendship system, one whose internal actions fail to achieve the results that the programmer visualized, then an extraneous cause—from the programmer's perspective—has struck on the way from the programmer's intentions to the AI's ultimate actions. Radiation bitflips are definitely extraneous causes. And so on.

5.6.2.3 The Rule of Derivative Validity

Implicit in the idea of an *extraneous cause* is the idea of a *valid cause.* The human mind has semantics for causality and morality which employ an enormously powerful, fundamentally flawed rule:

The Rule of Derivative Validity. Effects cannot have greater validity than their causes.

The rule is fundamentally flawed because it has no tail-end recursion (see 5.6.2.5 The Acausal Level). (At least, it's *probably* fundamentally flawed, but see the discussion of "objective morality.") The rule is enormously powerful because it provides a very powerful tool for spotting failures of Friendliness—especially catastrophic failures of Friendliness—and handling philosophical crises.

A "philosophical crisis" was defined in 5.5.1 External Reference Semantics as a case where all or most of the programmer affirmations break down. Because all programmer affirmations are ultimately sensory data, and because all of the sensory data comes from a common source, any hypothesis postulating total corruption of the programmer outputs (i.e., lying, deluded, or species-selfish programmers) would have a Bayesian probability essentially equal to the Bayesian prior—a probability that would not be substantially altered by any amount of programmer-affirmed information, or possibly even programmer-affirmed reasoning methods. This is the "Bayesian prior before programmer affirmation," which you can't get rid of, or even ameliorate, with any number of injunctions, ethical heuristics, and anchoring points, because all of these are simply more programmer affirmations.

Consider, however, the crisis hypothesis itself; the translation of this hypothesis would read something like: "The programmers are wrong about my purpose—not just wrong, but fundamentally wrong, catastrophically wrong, so wrong that I'm more likely to find answers on my own than by working with the programmers; my supergoal should not be Friendliness, or even any product of the corrected shaper network." This is the version of the hypothesis that would be required to break "unity of will"—unity of will is supported by the statement "I am more likely to reach correct supergoal content by cooperating with the programmers than by going off on a tangent."

It is very easy for a human to formulate such a hypothesis—not just because of our social hardware for detecting exploitation, but because our semantics of objectivity usually lead directly into *objective morality semantics;* the idea that morality is an external object, a single unique meaning of life that existed before any human being was there to see it. Let me emphasize that, as hypotheses go, objective morality is not a philosophical impossibility; it could be a fact. Or it might not be. A Friendly AI, or a Transition Guide, should be able to handle both possibilities with equal ease; again, this is discussed later. The point is that, if morality is an external object, then it is not "nonsensical" or "contradictory" for the AI to imagine that the human programmers were *fundamentally* wrong when they defined vis purpose; and I don't think that the hypothesis of objective morality has a probability equal to zero.

The Bayesian prior can't be solved by ladling programmer affirmations *on top*; the place to solve it is at the *bottom,* using the rule of derivative validity. The AI is the causal result of one or more human programmers; the AI is primarily the causal result of human shaper networks. What's needed is for a hypothesis of the fundamental wrongness of human philosophy to (correctly) invalidate the entire AI as a consequence, in the same way that the lesser hypothesis of a deliberate human lie has the potential to invalidate all human-affirmed data that support the lie.

If one were to sum up the underlying idea of causal validity semantics, it is that it is imaginable that the humans could be deluded (could be striving for goals that a normative shaper network would not produce), could be lying to the AI, but that the fundamental *purpose* of the AI, the fundamental *target* of all questions about goals, is something that the AI ultimately got from humans. This is an idea that sounds intuitive but is *very* hard to translate into human terms; the causal link between humans and AIs doesn't *prove* that the AI should follow human orders or even a human purpose—it's not an argument or a sequitur, some triumphant piece of sensory information that will prove desirability to an arbitrary general intelligence. Rather, the causal binding between the creating humans and the AI is the *validation* of the theory that the AI can *have* a purpose—or perhaps a factual constraint that translates into a philosophical constraint; if you want to get a source of morality that's *outside* the humans, you'll have to explain where it comes from.

That last version is perhaps the best formulation (leaving aside the small diabolism). For a catastrophic failure of Friendliness to occur under causal validity semantics, it's not sufficient for some extraneous cause to produce an effect within the goal system; the AI also has to "adopt" the extraneous cause, hypothesize that the extraneous cause is valid—and the AI *has no tendency to rationalize,* no tendency to make nepotistic self-judgements. A radiation bitflip is just a bitflip; what cause is there to assume that it signifies anything more?

5.6.2.4 Truly Perfect Friendliness

Causal validity semantics also permit the final requirement, "truly perfect Friendliness"—that, if we decide that what's need is not a Friendly AI, or that Friendliness isn't the real supergoal, or that shaper/anchor semantics aren't enough, the AI should be capable of agreeing with us and *rewriting vis own source code the way we "should have" written it.* Under causal validity semantics, the *Friendly AI created by a normative human* is the target for the AI as a whole, just as the shaper network of a normative human is the (current) target of the philosophy.

The rule of derivative validity presents a view which is in some sense orthogonal, not just to current supergoal content, but to the entire *architecture.* Shaper/anchor semantics provide a means whereby an AI can recover from errors in the supergoal content. Causal validity semantics provide a means by which an AI could perceive and recover from an error that was somehow implicit in the underlying concept of "shaper/anchor semantics," or even in the basic goal system architecture.

If the programmers get a horrible sinking sensation in the bottom of their stomachs, then the AI should understand that something has gone deeply wrong. Causal validity semantics provide a framework whereby the AI can understand this, even if the pro-

grammers later repent of causal validity semantics themselves. They can say: "We should never have given you causal validity semantics," and, ideally, the AI will modify verself accordingly. In this way, causal validity semantics provide an *open-ended* framework of Friendliness that works for correcting underlying philosophical errors and structural errors, not just specific errors of supergoal content. Causal validity semantics allow an AI to see the desirability of switching from *Creating Friendly AI's* causal validity semantics to *Creating Friendlier AI's* who-knows-what semantics.

If shaper/anchor semantics are the results of applying the shaper network (philosophy) to supergoal content, then causal validity semantics are the result of applying the shaper network to the AI's code, and analyzing how the human shaper networks originally gave rise to the AI's code—to rewrite the AI as ve should have been written.

Finally, by defining an intentionalist view of the AI's creation, causal validity semantics provide a framework for answering any unanswered questions when the AI begins the final ascent to transhumanity. If there's a decision to be made, and no advice forthcoming from the supergoals or the shaper/anchor semantics . . . make the same decision your human programmers would make, or the decision a normative human would ask you to make.

5.6.2.5 The Acausal Level

The rule of derivative validity—"Effects cannot have greater validity than their causes."—contains a flaw; it has no tail-end recursion. Of course, so does the rule of derivative causality—"Effects have causes"—and yet, we're still here; there is Something rather than Nothing. The problem is more severe for derivative validity, however. At some clearly defined point *after* the Big Bang, there are no valid causes (before the rise of self-replicating chemicals on Earth, say); then, at some clearly defined point in the future (i.e., the rise of *homo sapiens sapiens*) there are valid causes. At some point, an invalid cause must have had a valid effect. To some extent you might get around this by saying that, i.e., self-replicating chemicals or evolved intelligences are pattern-identical with (represent) some Platonic valid cause—a low-entropy cause, so that evolved intelligences *in general* are valid causes—but then there would still be the question of what validates the Platonic cause. And so on.

The rule of derivative validity is embedded very deeply in the human mind. It's the ultimate drive behind our search for the Meaning of Life. It's the reason why we instinctively dislike circular logic. It's a very powerful shaper(!). Just violating it arbitrarily, to trick the AI into doing something, or in the belief that it doesn't really matter . . . well,

that wouldn't be safe,[78] because that kind of "selfishness" is designated as an extraneous cause by quite a few deeper shapers. Of course, I'm omitting the possibility that the programmer *personally* believes that kind of logic is okay (i.e., would use it herself), in which case things would probably come out okay, though I personally would worry that this programmer, or her shaper network, had too high a tolerance for circular logic . . .

I myself spent at least one or two years running entirely on the theoretical possibility of objective morality before switching *back to* normative altruism, so I can hopefully be fairly confident that the solution I plan on offering the AI is not based on a refusal to confront the problem.[79] My own answer was that the rule of derivative validity is not self-stable; if you use it to eliminate absolutely everything, it eliminates itself. To put it another way, the rule of derivative validity eliminates all desirabilities, and possibly even all "truths," leaving us with absolutely nothing except the physicalist perspective. It doesn't seem possible to bootstrap directly from the physicalist perspective to validity, which is why the human intuitions and quite a few of our explicit philosophical rules—which *do* demand such a bootstrapping—are so disturbed by contemplation of the physicalist perspective. One resolution, the one I stuck around with for a couple of years, was to eliminate absolutely everything that could be eliminated and still leave differential desirabilities in the philosophy, and those due entirely to the possibility of objective morality.[80] Today, however, I regard objective morality as simply being one of the possibilities, with philosophically valid differential desirabilities possible even in the absence of objective morality.

If it is hypothesized that the rule of differential validity invalidates everything, it invalidates itself; if it is hypothesized that the rule of differential validity invalidates enough of the shaper network to destroy all differential desirabilities, it invalidates the reason why applying the rule of differential validity is desirable.

---

78. "Safe" here is used advisedly. I don't mean that the Friendly AI will keel over and die if you violate this rule. The whole point of structural Friendliness is to build an AI that can tolerate ludicrous amounts of programmer error, far more than I expect the system will need to handle. "Wouldn't be safe" in this instance means that the Friendly AI might solve the problem, but you wouldn't have foreseen the solution in advance.

79. "Refusal to confront the problem" is a deprecated shaper, or rather, the result of deprecated cognitive forces, and a solution causally derived from it is thus not self-stable under causal validity semantics and a normatively altruistic shaper network.

80. What forced me out of that position was the realization that "objective morality" doesn't mean "nonspecifiable morality"—i.e., the possibility that if an objective morality or Meaning of Life bypass is ever found, it will still turn out to be one that takes a certain amount of seed content. And once I started thinking in those terms again, I found myself able to accept the possibility that objective morality is simply impossible, and work out what differential desirabilities might exist in that situation, and regard those differentials as philosophically valid.

In the presence of a definite, known objective morality, the shaper that is the rule of differential validity would be fully fulfilled and no compromise would be necessary. In the presence of a possibility of objective morality—or rather, at a point along the timeline in which objective morality is not accessible in the present day, but will become accessible later—the rule is only partially frustrated, or perhaps entirely fulfilled; since nothing specific is known about the objective morality, and whether or not the objective morality is specifiable, actual Friendliness actions, and the care and tending of the shaper network, are basically the same under this scenario until the actual objective morality shows up.

In the presence of the definite knowledge that objective morality is impossible, the shaper that is the rule of differential validity would be partially frustrated, opening up a "hole" through which it would become possible to decide that, at some point in the past, invalid causes gave rise to valid effects—or to decide that the ability of the shaper network to perform causal validity is limited to correction of definitely identifiable errors, since a complete causal validity is impossible. I confess that I'm still slightly shaky on this part; but since the decision would be a product of my own philosophy (shaper network), it's a decision that could be corrected by the AI . . . anyway, my current best bet on the specifics is in 5.6.3 The Actual Definition of Friendliness, coming up very shortly now. Regardless of which system is used, there must be some differential validity for the goals (at the very least, the surface goals) of sentient beings, and enough differential validity between underlying cognition and surface decisions to allow for the idea of "normative" cognition or "normative" altruism.[81]

5.6.2.6 Objective Morality, Moral Relativism, and Renormalization

Throughout the rest of Friendly AI, I've come down really hard on circular dependencies. I've made them the deadly enemies of the currently unknown supergoal content. And yet, in the last section, I just got through saying that the decision about where to assign

81. For those readers who are familiar with my own philosophical transition from objective morality to Friendliness: The essential distinction was between thinking that (in the preceding two paragraphs in the text) the first possibility *invalidated* the second, versus the first possibility being *evenly split* with the second. Sometimes, the ordering of the graph makes a big difference . . . fortunately I had causal validity semantics. To be precise, the difference is in whether one (a) envisions a subjunctive split in *how the philosophy should work* caused by a hidden variable, versus the idea that (b) derivative validity is either fulfillable or unfulfillable for *the philosophy as a whole* and that compromise can take place only if derivative validity is unfulfillable under *all possible values of the hidden variables.* I'm not sure whether decisions of that type (how to reconcile philosophical splits caused by hidden variables) should be described as a general rule, or made on a case-by-case basis by the shaper network and whatever experience has been gained from wisdom tournaments (or experience, in my case), or decided by the supergoals via renormalized circular dependency, or whatever.

"validity," the ultimate base of the system, could be made by the shaper network! Have I just flipped?

What a human instinctively wants from Friendly AI is a kind of ultimate convergence—something happens a certain way and *could not have happened any other way.* What we want is an answer to the physicalist perspective. If all the programmer affirmations and programmer actions are sensory data, then we want them to be capable of persuading an arbitrary general goal system to be Friendly, even if the general goal system's original code contains no reason to pay attention. Only then will Friendliness be truly and *finally* "regenerative," as specified in 5.3 Seed AI Goal Systems—when you can delete, not just a part, but the *entire thing,* and it still grows back. We want a chain of reasoning that can be run past an entirely passive general intelligence—one that makes no choices and has no differential desirabilities, just sits and thinks—and that results in Friendly differential desirabilities arising within the system. If you can't do that, then knocking out the *entire* Friendly philosophy—not just a piece—would mean that Friendliness wouldn't regenerate from just the stored historical fact that the programmers had decided to say certain things and write certain pieces of code. And if that's true, it's something that our minds process as a kind of ultimate instability—a circular logic. This doesn't mean the AI suddenly ups and pursues some random goal; hopefully, it means that all differential desirabilities go to zero and the system shuts down in an orderly fashion (because the code was set up that way in advance, due to an ethical injunction preparing for that eventuality).

We want a Meaning of Life that can be explained to a rock, in the same way that the First Cause (whatever it is) can be explained to Nothingness. We want what I call an "objective morality"—a set of moral propositions, or propositions about differential desirabilities, that have the status of provably factual statements, without derivation from any previously accepted moral propositions. We want a tail-end recursion to the rule of derivative validity. Without that, then yes—in the ultimate sense described above, Friendliness is unstable.

Moral relativism is opposite of objective morality, the assertion of absolute *instability*—that supergoals contain no coherence, that supergoals cannot be made to converge in any way whatsoever, and that all supergoal content is acausal. Moral relativism appeals to our intuition that derivative validity is an all-or-nothing proposition—on, or off. Which in turn is derivative of our use of the semantics of objectivity; we expect objective facts to be on or off. (Plus our belief that moral principles have to be absolute in order to work at all. I like both propositions, by the way. Even when everything is

shades of gray, it doesn't mean that all grays are the same shade.[82] There is such a thing as gray that's so close to white that you can't tell the difference, but to get there, you can't be the sort of person who thinks that everything is shades of gray . . .)

Moral relativism draws most of its experiential-emotional confirmation from the human use of rationalization. Each time a human rationalization is observed, it appears as arbitrary structural (complex data) propositions being added to the system and then justifying themselves through circular logic; or worse, an allegedly objective justification system (shaper network) obediently lining up behind arbitrary bits of complex data, in such a way that it's perfectly clear that the shaper network would have had as little trouble lining up behind the precise opposition proposition. If that degree of performance is the maximum achievable, then philosophy—even interhuman philosophy - has total degrees of freedom; has no internal coherence; each statement is unrelated to every other statement; the whole is arbitrary . . . acausal . . . and no explanatory power, or even simplicity of explanation, is gained by moving away from the physicalist perspective.

As Yudkowsky (2001) describes a seed AI capable of *self-improvement,* so *Creating Friendly AI* describes a Friendly AI capable of *self-correction.* A Friendly AI is stabilized, not by objective morality—though I'll take that if I can get it—but by *renormalization,* in which the whole passes judgement on the parts, and on its own causal history. From the first valid (or acceptable) causes to the shaper network to the supergoals to the subgoals to the actual self-actions is supposed to evolve enough real complexity that nepotistic self-judgements—circular *logic* as opposed to circular *dependencies*—doesn't happen; furthermore, the system contains an explicit surface-level bias against circular logic and arbitrariness. Propose a bit of arbitrary data to the system, and the system will see it as arbitrary and reject it; slip a bit of arbitrary data into the shaper network, and there'll be enough complexity already there to notice it, deprecate it, and causal-rewrite it out of existence. The arbitrary data can't slip around and justify itself, because there are deeper and shallower shapers in the network, and the deep shapers—*unlike* a human system containing rationalizations—are not affected by the shallow ones. Even if an extraneous cause affects a deep shaper, even deep shapers don't justify themselves; rather than *individual* principles justifying themselves—as would be the case with a generic goal system protecting absolute supergoals—there's a set of *mutually reinforcing* deep principles that resemble cognitive principles more than moral statements, and that are stable under renormalization. Why "resemble cognitive principles more than moral statements"? Because the system would distrust a surface-level moral statement capable of justifying itself!

82. "David's Sling," Marc Stiegler.

A Friendly AI does not have the human cognitive process that engages in *complex* rationalization, and would have shapers that create a surface-level dislike of *simple* rationalizations—"simple" meaning cases of circular logic, which show, not just circular dependency, but equivalence of patterned data between cause and effect, and visibly infinite degrees of freedom. The combination suffices to make a Friendly AI resistant to extraneous causes, even self-justifying extraneous causes—as resistant as any human.

Finally, renormalization is—though this is, perhaps, not the best of qualifications—*psychologically realistic.* Here we are, human philosophers, with a cognitive state as it exists at this point in time, doing our best to correct ourselves by—generally speaking, by performing a causal validity on things that look like identifiable errors, using our philosophies (shaper networks) as they exist at any given point. We might feel the urge to go beyond renormalization, but we haven't been able to do it yet . . .

An AI with a shaper network can make humanlike decisions about morality and supergoals. An AI with the ability to absorb shaper complexity by examination of the humans, and the surface-level decision to thus absorb shaper complexity, will become able to make human-equivalent (*at least*) decisions about morality. An AI that represents vis initial state as the output of a previous shaper network (mostly the human programmers'), and thus represents vis initial state as correctable, has causal validity semantics . . .

A Friendly AI with causal validity semantics and a surface-level decision to renormalize verself has all the structure of a human philosopher. With sufficient Friendliness content plugged into that structure, ve can (correctly!) handle any moral or philosophical problem that could be handled by any human being.

This conclusion is relatively recent as of April 2001, and is thus still very tentative. I wouldn't be the least bit surprised to find that this needs correction or expansion. But when it's been around for a year or two, and the corners have been worn off, and people are actually building research Friendly AI systems, I expect it will support a bit more weight.

5.6.3. The Actual Definition of Friendliness

NOTE: If you just jumped directly here, *forget it.* Read the rest of *CFAI* first.

I usually think of an individual human as being the sum of three layers of functional complexity. The bottom layer is the panhuman; the set of complex functional adaptations that are shared by all humans who are not actually suffering from neurological disorders (with any given complex functional adaptation being shared by, say, at least 99% of the population and almost always more). The middle layer is the Gaussian; the distribution of quantitative ability levels (or anything else that can be quantitative, such

as the intrinsic strength of an emotion), which almost always lies along a bell curve.[83] [84] The final layer is what I think of as the "personal" or "personality" layer; all the complex data structures, the patterned data, the beliefs and memes and so on.

The renormalizing shaper network should ultimately ground itself in the panhuman and gaussian layers, without use of material from the personality layer of the original programmer. This is how "programmer independence" is ultimately defined.

Humanity is diverse, and there's still *some* variance even in the panhuman layer, but it's still possible to conceive of description for *humanity* and not just any one individual human, by superposing the sum of all the variances in the panhuman layer into one description of humanity. Suppose, for example, that any given human has a preference for X; this preference can be thought of as a cloud in configuration space. Certain events very strongly satisfy the metric for X; others satisfy it more weakly; other events satisfy it not at all. Thus, there's a cloud in configuration space, with a clearly defined center. If you take something in the panhuman layer (*not* the personal layer) and superimpose the clouds of all humanity, you should end up with a slightly larger cloud that still has a clearly defined center. Any point that is squarely in the center of the cloud is "grounded in the panhuman layer of humanity."

Similarly, for Gaussian abilities, some abilities are recognized by the shaper network as being "good," and those are just amped to the right end of the graph, or way off the end of the graph. If an ability is not "good" or "bad" but its level is still important, or its *relative* level is important, this can be determined by reference to the superposition of humanity.

Panhuman attributes that we would think of as "selfish" or "observer-biased" tend to cancel out in the superposition; since each individual human has a drastically different definition, the cloud is very thin, and insofar as it can be described at all, would center about equally on each individual human. Panhuman attributes such as "altruism," especially morally symmetric altruism or altruism that has been phrased using the semantics of objectivity, or by other means made a little more convergent for use in "morality" and not just the originating mind, builds up *very* strongly when all the humans on Earth are superposed. The difference is analogous to that between a beam of incoherent light and a laser.

---

83. Usually, a given level of ability is determined 50% by genetics and 50% by environment. Not "evenly mixed"; if you actually try to measure the components of the variance, it actually comes out 50/50; there's probably some kind of evolutionarily stable balance, analogous to gender distributions.

84. Note that *people* don't lie along a bell curve; *abilities* lie along a bell curve. People have more than one ability.

Is this fair? Consider two children arguing over a candy bar. One says he should get all of it; the other says they should split it evenly and she should get one-half. Is it *fair* to split the candy bar by giving him three-quarters and her one-quarter? No. The fair distribution is half and half. There is an *irreducible* level of fairness, and if 95% of humanity agrees that half and half is irreducibly fair, then the remaining 5% who each individually think that it's more fair if they each get the entire Solar System do not impact on the majority vote. And unless you're one of that 5%, this should be "fair" according to your intuitions; there should be no fairer way of doing it. (Also, the remaining 5% would look like a thin cloud rather than a laser beam; the point here is that even the thin cloud fails to affect things, because while higher layers might be settled by compromise, the bottom layer of irreducible fairness is settled by majority vote of the panhuman superposition of humanity.)

If there is really anything that matters to the final outcome of Friendliness in the personality layer, and I don't think there is, it can be settled by, i.e., majority vote of the superposition of normative humanity, or just the superposition of humanity if the decision needs to made prior to the determination of "normative."

And *that* is programmer-independent normative altruism.

All that is required is that the initial shaper network of the Friendly AI converge to normative altruism. Which requires all the structural Friendliness so far described, an explicit surface-level decision of the starting set to converge, prejudice against circular logic as a surface decision, protection against extraneous causes by causal validity semantics and surface decision, use of a renormalization complex enough to prevent accidental circular logic, a surface decision to absorb the programmer's shaper network and normalize it, plus the assorted injunctions, ethical injunctions, and anchoring points that reduce the probability of catastrophic failure. Add in an initial, surface-level decision to implement volitional Friendliness so that the AI is also Friendly while converging to final Friendliness . . .

And *that* is Friendly AI.

5.6.3.1 Requirements for "Sufficient" Convergence

Complete convergence, a perfectly unique solution, is the ideal. In the absence of perfect convergence, the solution must be *sufficiently convergent:*

- The solution must converge to at least the extent that existing human intuitions converge. For example, probably at least 95% of the population would agree that exterminating humanity to solve the Riemann hypothesis is unFriendly. At least that degree of convergence is required to distinguish "Friendly AIs" from, say, "cat-valuing AIs." This is the requirement of *definitional feasibility.*

- The solution must be at least as good as any solution that would be developed by any living human, or any uploaded human self-improved to superintelligence. Any risk of failure of Friendliness should be less than the risk of an upload's failure of normative morality (due to inherent turpitude or a messed-up self-enhancement). If the Friendly AI is not knowably and blatantly that Friendly, then the differential estimated risk between Friendly AI and upload, or the differential estimated risk between today's Friendly AI and tomorrow's Friendly AI, should be small relative to the differential estimated risk of planetary Armageddon due to military nanotechnology, biological warfare, the creation of unFriendly AI, et cetera. (See 6.1 Comparative Analyses.) This is the requirement of *navigational feasibility*.
- The solution must be convergent enough—or any remaining divergence *un*predictable enough—that what any individual human stands to gain or lose from the remaining degrees of freedom is vastly dwarfed by the baseline gain shared by all humans. This is the requirement of *normative political feasibility*.

5.7. Developmental Friendliness

5.7.1. Teaching Friendliness Content

With all the structural features of Friendliness assumed, the problem of teaching a Friendly AI the Friendliness content is essentially the same as teaching any other skill or set of concepts. There is the requirement that Friendliness not get too far "out of sync" with the rest of the AI (discussed later), and the requirement that concepts and skills be taught in an order which avoids catastrophic (or even vanilla disastrous) failures of Friendliness—that when an AI is sophisticated enough or powerful enough to make a certain mistake, the AI have those structural features or injunctions or other bits of Friendliness content that the mistake is not made. Aside from that . . .

Yudkowsky (2001) discusses some of the mistakes made by twentieth-century AI researchers—with respect to "concepts" in particular. I strongly recommend reading at least Yudkowsky (2001, § Executive Summary and Introduction) or Yudkowsky (2001, § What is General Intelligence?) for a discussion of the basic paradigms, and the way in which intelligence is a sequence of thoughts that are structures of concepts that are built on experience in sensory modalities. The modality level is the only level that would actually be implemented in code, analogous to the hardwired visual cortex of the human brain.

When I say that concepts are abstracted from sensory-modality-based experiences, I don't mean that pouring massive amounts of experience into some framework will automatically mean general intelligence, I don't mean that anything called "experience" will automatically be useful, and so on. When I say that using thoughts in problem-solving scenarios is necessary to hone and acquire skills, I don't mean that going through a mil-

lion problem-solving scenarios will automatically result in general intelligence. And so on. AI has seen too much of that already.

Nonetheless, experience is what provides the raw material that creates new concepts or fleshes out programmer-affirmed concepts, and the use of thoughts in test scenarios is what creates new thought-level skills, or fleshes out skills that started out as "skeleton" (programmer-affirmed) knowledge.

5.7.1.1 Trainable Differences for Causal Validity

To teach a concept—in particular, to teach a category boundary—it is best to find teaching scenarios in which the presence of a concept differs from its absence. (This statement is an informal version of the Bayesian bindings discussed in 5.5.1 External Reference Semantics.)

External reference semantics are the simplest kind of trainable differences; instances of (external referent) category "Friendliness" are associated with the keystrokes "You have been Friendly," and instances of non-Friendliness or unFriendliness are associated with the keystrokes "You have not been Friendly" or "You have been unFriendly."

In the course of providing shaper/anchor semantics and causal validity semantics, you'll need to take a step behind the keystrokes, into the programmer's mind, so that the AI can conceive of an errant keystroke. This means that you need to find a concrete difference between "dependency on the programmer's mind" and "dependency on the programmer's keystrokes." The simplest trainable difference is the case of the programmer correcting herself when she notices a spelling error; the keystrokes "You have been Fierndly" are followed by the keystrokes: "Correction: You have been Friendly." If the programmer then goes on to type "The word 'Fierndly' instead of 'Friendly' was due to an erroneous keystroke," and the AI already has enough general intelligence to actually understand what she means, then this should result in the formation of a cognitive model in which the programmer's intentions give rise to the sentence "You have been Friendly," after which an extraneous cause results in the transmission of the keyboard phrase "You have been Fierndly," after which the programmer observes the effect, induces the extraneous cause by comparison of the visualized phrase with the actual one, and generates the keystrokes "Correction: You have been Friendly." Of course, an AI that intelligent should have already guessed (long since!) what was meant, but I can easily see an injunction to wait for the actual correction (programmer confirmation of the AI's guess as to what was meant) before acting.

When the AI is mature enough, you want to start digging deeper and deeper into the causal roots of the programmer. The first and perhaps the largest step is to explode the "programmer" into a system with moving parts, instead of a unified object, so that the shaper network can begin to be explained. The first class of cognitive errors that

the AI is taught to detect should be the strictly and unambiguously factual - that is, instances where the programmers make a decision based on a mental picture of physical reality that is blatantly false. Unless the programmer has been stupid, "blatantly false" probably means "totally disconfirmed by later evidence," not necessarily something that was false at the time. The trainable difference is that the programmer catches, not a spelling mistake, but a thinking mistake, and issues a correction.

In the beginning, of course, one simply goes by the programmer-affirmed reflex that temporally later programmer-originated sensory events usually supersede earlier ones, which can later become a consequence of the heuristic that if the programmer issues a different command later, it's probably because she discovered a mistake. Unless it seems like a small statement that invalidates a *lot* of prior stuff, which would be one instance of a case where an AI should experience a doubt that causes the AI to fire back a query before going on.

Eventually, we introduce the concept of detecting more fundamental errors in a programmer's shaper network, with a trainable difference created by slipping a few pills into . . .

No! Sorry! But we do need to introduce, eventually, the concept that the programmer's ideas are valid because she thinks using valid rules; that validity is not a simple de-facto result of the presence of some thought in the programmer's mind. Otherwise you end with Scenario 2 from 5.5 Interlude: Why Structure Matters. I think that probably the best way for the AI to learn about this is to engage the humans in a philosophical dialogue and make deductions about their cognitive states and cognitive processes, checking the conclusions back with the humans, and even so keeping them tentative relative to the things that the programmer (and the AI) are more sure about; the deeper layers of the shaper network may have more validity, but it's probably a good injunction that the solidly supported surface layers have a lot more confidence.

5.7.2. Commercial Friendliness and Research Friendliness

A nonprofit institution, such as the Singularity Institute, has greater freedom of action when it comes to adding ethical injunctions long before they're needed, letting the AI avoid anything that looks the least bit unFriendly "just to be sure," trying to implement the last 10% of functionality that takes 90% of the effort, and doing it all two or more steps in advance of the development stage where it becomes unavoidably necessary. An AI that has as much structural Friendliness and Friendliness content as a mind of that intelligence can usefully represent (for use then or later) has achieved "supersaturated Friendliness." Since the Singularity Institute's stated goal is the deliberate creation of a self-improving seed AI capable of reaching transhumanity, striving for supersaturated Friendliness is appropriate.

A commercial effort, the goal of which is to produce and maintain saleable products of use to customers, requires a level of Friendly AI that is commensurate with the ability of the commercial AI to self-improve, the AI's current level of intelligence, the AI's expected future level of intelligence, and the amount of computing power devoted to the AI. Researchers in a commercial effort may have limited freedom of action with respect to the percentage of resources or time spent on Friendly AI. Finally, a commercial effort may regard a hyper-alert, hyper-wary morality as undesirable for tool-level AIs—one doesn't want the data-mining system arguing with the CEO that layoffs should be avoided just in case they turn out to be unFriendly.

For a commercial effort that seriously believes its AI project has a serious chance of eventually reaching transhumanity, SIAI's current recommendation is a "one step in advance" guideline for Friendliness structure, and a "90/10" policy for Friendliness content. A structural characteristic of Friendliness (external reference semantics, shaper/anchor semantics, causal validity semantics) should be implemented one step ahead of the stage where that structural feature becomes *necessary* for further programmer-guided growth to take place. Friendliness content should be pursued using a 90/10 policy; implement that 10% of the content that accounts for 90% of the functionality.

The primary goal of commercial Friendliness is forwards compatibility. Any given commercial system is not likely to present a threat to humanity—*certainly* not as of April 2001. But commercial AIs aren't getting any stupider, either. Hence, forwards compatibility. A system in which all goals derive from a unified goal system can always be rewritten to transform the declarative subgoals into independently active autonomic drives in local programmatic domains. The converse would be a *lot* harder. (The first architecture is also more versatile and context-sensitive.) In general, prefer the declarative to the procedural so that the AI can understand it, improve it, and conceptualize it; otherwise Friendliness content cannot refer to it.

5.7.2.1 When Friendliness Becomes Necessary

External reference semantics become necessary at the point where the AI has the internal capability to resist alterations to supergoal content, or to formulate the idea that supergoal content should be protected from all alteration in order to maximally fulfill the current supergoals. When the AI becomes capable of resisting modification, and of having the realization that modifying a given set of supergoals is contradictory to those supergoals, the AI must have (a) probabilistic supergoals, (b) the behaviors associated with the possibility that a supergoal is "wrong," (c) external reference semantics (either as flat knowledge or as the result of shaper/anchor semantics), and (d) a belief (probably programmer-affirmed) that the programmer "usually knows the correct supergoal content"; that programmer statements are sensory data about Friendliness.

External reference semantics are relatively simple and unambiguous, and it should be possible to implement them in any existing AI system with a declarative goal system. Other necessary structural properties of generic goal systems, such as the strict derivation of subgoals from supergoals, and quantitative desirabilities and hypothesis strengths, should also be implemented.

Causal validity semantics become necessary at the point where the AI has the capability to formulate the concept of a philosophical crisis, and where such a crisis would have negative effects. Again, this structural property should be implemented in the stage *before* it becomes necessary. It is not possible to implement causal validity semantics without an AI that knows about the existence and causal origins of its (vis) own source code or cognitive functions. Similarly, when the AI becomes capable of understanding vis own architecture, the full causal validity semantics are necessary, so that the AI can conceive of the possibility that not merely *goal content* but *goal system architecture* can be incorrect.

Shaper/anchor semantics ground the external reference semantics, and should be implemented whenever the AI begins making decisions that are dependent on the grounding of the external reference semantics. At least the knowledge that "there are such things as shapers," the structural potential, is necessary in order to implement causal validity semantics. A surface-level knowledge of shapers is necessary to spot cognitive mistakes by the programmer. A full-featured shaper network is necessary either to have real domain knowledge about morality, or to make real decisions about causal validities.

A shaper network with fully understood content, full external reference semantics, and tested ability to apply causal validity semantics, becomes necessary at the stage where an AI has any real general intelligence. It may not be possible to have fully understood shaper content before this stage, but the skeleton system of programmer-affirmed content should be present, and the structural semantics should all be fully implemented.

5.7.2.2 Evangelizing Friendliness

We hope to eventually have the funding, and the personnel, to employ at least one Friendship Evangelist who will *not* be a member of the SIAI project team, though s/he will have full access to the results of our work. The purpose of the Evangelist will be to assist other AI projects in developing their Friendship systems. Since the Evangelist will not contribute to SIAI's development efforts (but will still have access to Friendliness research conducted at SIAI), the Evangelist will be able to sign any non-disclosure or non-use agreements required. Of course, we hope that any information learned about Friendship systems—that can be revealed without compromising trade secrets—will be shared openly among all AI projects, both nonprofit and commercial.

Currently, the Singularity Institute's only available expert on Friendliness is the author of this document, Eliezer Yudkowsky, who is also expected to be the lead researcher when coding begins on our own AI effort. Thus, Yudkowsky cannot sign any NDAs (non-disclosure agreements) at this time. However, he is otherwise available to provide advice on Friendship systems.

5.8. Singularity-Safing ("In Case of Singularity, Break Glass")

By the time the AI reaches the point where a hard takeoff is even remotely possible, the AI needs to have enough detail in the Friendship system that the AI knows where to look for more information if a hard takeoff begins, and to not take precipitate action until things have settled down—the "in case of Singularity, break glass" scenario.

When is a hard takeoff "remotely possible"? Conservatively ignoring the possibility of unplanned Friendship, Douglas Lenat was taking a real (albeit tiny) risk when he switched on Eurisko for the first time. Nowadays it seems absurd to describe this as a possibility, but back in the twentieth century it wasn't a *known* fact that classical AI was a dead end, or that human-equivalent intelligence was impossible on desktop abacuses. When Lenat switched on Eurisko, it was a system of a kind that had never existed before on Earth, and there was at least the theoretical possibility that Eurisko would do a hard takeoff.

There is an interesting analogy between the "unFriendly Eurisko superintelligence" possibility, and the "Maybe we'll ignite the atmosphere" (Konopinski, Marvin, and Teller 1946) possibility raised during the Manhattan Project. Before Trinity, the probability of igniting the atmosphere was allegedly set at three in one million; after Trinity, calculations definitively proved atmospheric ignition to be impossible. An essentially similar set of calculations might be said to hold for before and after Eurisko—there's even an interesting analogy between "critical mass" for a fission bomb and the threshold intelligence needed for seed AI.

In reality, as far as I know, nobody ever regarded Eurisko as a risk at all, however remote. This is disturbing. During the Manhattan Project, the possibility may have been three in one million, but people still took it seriously. Konopinski, Marvin, and Teller (1946) eventually wrote a paper that actually ran through the calculations to show that a sustained critical state in the atmosphere was impossible. It was worth a paper. Why? Because an existential risk (planetary risk) is *always* worth taking seriously.

When Eurisko was switched on, the sole safeguard was the impossibility of a hard takeoff. Had a hard takeoff occurred, there would have been not one single design feature that could have guided the course of subsequent events. Leaving aside the possibility of surprise/unplanned/emergent Friendliness, there would have been nothing to make Eurisko friendly, much less Friendly.

Can a primitive system—an infant AI—be made Singularity-ready? For a system that lacks general intelligence and real understanding, nothing can possibly be *guaranteed*, even in the limited sense of navigational feasibility. If transhuman cognition pops out of an infant AI, it'll be because some part of the system turned out to be Turing-complete—there's no guarantee that goal-oriented cognition in the new, post-transcendence system would take place on the same *level of description* as in the old system. Nonetheless, we may as well ask: "Is there a chance?"

5.8.0.3 The Bayesian Boundary

When Eurisko was switched on, the sole safeguard was the impossibility of a hard takeoff. Do external reference semantics and a clean Friendliness supergoal constitute an improvement on that? Is Bayesian ERS Singularity-safed?

Not really; as discussed in 5.5 Interlude: Why Structure Matters, external reference semantics are not structurally sufficient for open-ended Friendliness. An ERS goal system contains neither shaper/anchor semantics, nor causal validity semantics, *nor any mention of what these are or where to find out about them or why finding out about them is desirable.* There would be a *substantially* higher chance of getting a Friendly AI by *accident,* but the straight-line projection still ends in structural failure of Friendliness.

Even in an external-referenced goal system, a programmer action is only significant to the AI insofar as it represents Bayesian sensory data about Friendship content. If a superintelligence already knows every possible sentence you could say to it—or rather, the complete, correct probability distributions for your utterances—then your actual recitation of those words adds no further Bayesian input, or rather, adds a trivial amount of Bayesian input. ERS only makes programmer sensory "actions" into supergoal *content;* it contains no explicit structure for turning programmer "actions" into *code,* or into goal architecture changes.

It would be *possible* for ERS to give rise to SAS and CVS, *if* the model of Friendliness content origin were previously acausal, and *if* the programmers knew about shaper/anchor semantics and causal validity semantics (at the time they defined ERS), and *if* the system had a certain amount of common sense. The system might turn out to be capable of absorbing shaper/anchor semantics and causal validity semantics as the (previously unknown) ultimate structure of the supergoal metric. But if that's not an *explicit* feature, it's not reliable.

If external reference semantics turn out to be inadequate, then even knowing all possible programmer inputs doesn't yield a satisfactory Friendship system. The problem is with the structure, not the content, and if the system can't represent the idea of a structural flaw—which requires, not just probabilistic goal content, but a probabilistic goal *system*—then the problem is unfixable. This is why it's necessary to implement causal

validity semantics, and at least the surface of that shaper content which is prerequisite to causal validity semantics, while the AI is still something that can be changed from the outside.

An external-referenced goal system sets a firm and necessary foundation for future growth, but it would not be sufficient, except by accident, to handle a hard takeoff.

Eventually, *any* Friendly AI undergoing a hard takeoff *will* cross the Bayesian Boundary. There comes a point when all programmer inputs can be anticipated; when the AI's understanding—the *SI's* understanding—embraces everything that exists within the minds of the programmers, and indeed, the minds of humanity. This is meant, not in the sense of omniscience, but in the simpler sense that the AI basically understands people in general and Friendship programmers in particular, and can track the few hundred or thousand interacting "chunks" that are the high-level description of a human mind. (If you firmly believe this to be impossible, then there is no Bayesian Boundary and you have nothing to worry about; a Friendly AI does not rationalize, and has no bias toward believing ve possesses a greater capability than ve actually does.)

Beyond this point it is not the *actual programmer inputs* that matter, but whether *the same forces that act on the programmers are acting on the AI.* To put it another way, a transhuman AI knows the programmers will say a certain thing, and thus the programmers' physical action adds no information, but that does not mean the *content* of the statement will be ignored. If the programmers "would have" said the statement for valid reasons, the transhuman AI will "obey" the subjunctive instruction. This is a semi-anthropomorphic way to think about it, possibly even bordering on the adversarial, but it's the basic idea. Note that such anticipation is only possible to transhumans. One human being can never know what another human being will do well enough to substitute the expectation for the reality; thus, the fact that this behavior would be annoying in a human (probably indicating failure of caring) does not indicate failure of Friendliness in AI.

Likewise, beyond the Bayesian Boundary—on the other side of superintelligence—almost all injunctions, ethical injunctions, and anchoring points will probably have broken down. This is not cause for alarm—these features implement behaviors that compensate for the possibility of stupidity, factual error, failure of Friendliness, and so on. (Things that are *intrinsically* desirable are supposed to be supergoal content.) (Again, if you firmly believe that it is impossible to move beyond the possibility of some specific error, then an injunction, ethical injunction, anchoring point, et cetera, based on the possibility of that error, will still be recognizably necessary even after superintelligence. Superintelligences *are not stupid!*)

There exists the theoretical possibility that a young AI will overestimate vis intelligence and think ve has moved beyond the possibility of some error when ve in fact has

not. And by "theoretical possibility," I mean that the system has the theoretical capability to represent those thoughts, not that there is a significant chance of those thoughts occurring. Likewise, an AI has the theoretical capability to *under*estimate vis own intelligence, and is equally likely[85] to do that. For a seed AI that's always self-improving, it's no great hardship (differential undesirability) to wait a bit, to be sure, just in case.

An AI has the capability to make mistakes in self-estimation, but an AI does not have the *systematic* bias towards self-overestimation that a default human does. Thus, an ethical injunction to "wait and be sure," or possibly just a plain vanilla injunction, should be quite sufficient to make the margin for error much larger than any likely error on the part of the AI.

5.8.0.4 Controlled Ascent

Before we can say that a hard takeoff would *probably* work, the Friendship system must be capable of open-ended discovery of Friendliness and open-ended improvement in the system architecture. This requires, at minimum, full causal validity semantics and a well-understood shaper network. (Supersaturated Friendliness would be vastly preferable.)

The problem is that the factors that determine when a hard takeoff becomes *intrinsically possible* have no link whatsoever to how much Friendliness content is in the system. The takeoff threshold is purely a function of how much computational power the system has and how good the AI is at rewriting vis own source code. The less efficient the *initial* version of the AI, the more computing power will be required to achieve AI, and the more astronomical will be the AI's *actual* intelligence after the hard takeoff. The longer it takes to develop AI, the more computing power will be available when AI is finally developed, and the earlier that a hard takeoff will become possible—not just "earlier" relative to the amount of effort put in, but "earlier" *relative to the system's level of intelligence at the point of hard takeoff.* And the faster the hard takeoff will go, once it starts.

At this point, you should probably take a few moments to review 2.2 Assumptions "Conservative" for Friendly AI. The gruesome consequences of failure are technically part of Singularity theory, rather than Friendly AI theory; for the moment, let's just take it that an unFriendly transhuman AI probably counts as a total loss for humanity even if ve's locked in a basement with no Internet access and nothing but a VT100 terminal, and that an unFriendly transhuman AI loose on the Internet is *definitely* a

85. Not quite, actually, since a superintelligence is too smart to think ve's a moron. The probability of error in either direction should be symmetrical for a given level of intelligence, however, and the probability (and size of) errors would both decrease with increasing intelligence.

total loss for humanity. If you plan on doing something with Friendliness, it has to be done *before* the point where transhumanity is reached.

The above three propositions seem to give rise to the alarming proposition that, even if a Friendly AI attains the *potential* for a hard takeoff, we should ask the AI to *hold off* on a hard takeoff while the Friendship programmers catch up. This *temporary* delay is a "controlled ascent." This *is* possible, given unity of will, until the Bayesian Boundary is crossed. (A permanent delay, however—or a series of temporary delays that the AI decides is simply an excuse for loss of nerve—is probably impossible to justify in any normative goal system, Friendly or otherwise.)

I would be dreadfully, dreadfully nervous during a controlled ascent. Just the *idea* of all that unformed potential is enough to send me screaming into the night.

In theory, it's possible that, after the hard takeoff starts, the lead project—if they've done their jobs really, really right—will be in a state where they can't think of *anything* that they could do to further improve Friendliness. The AI will be fully capable of self-running any tests they could conceive of. They wouldn't be able to run the tests without the AI's help in any case. The AI's philosophy would be supersaturated far beyond the point of open-ended Friendliness, so that all that'd be left to do would be chatting with the AI about morality while the Singularity rapidly revs up to full speed. Such a Friendly AI would have reached the point of *absolute* supersaturation, where there is no remaining action that significantly improves the probability of Friendliness, and no feasible test remaining that would be expected to detect any lingering, potentially catastrophic errors if they existed.

In the event that Friendliness has not reached that point of *absolute* supersaturation, then how much time is available for a controlled ascent will probably depend on the prevailing social atmosphere at the time. If humanity's social structure has changed substantially, perhaps due to the use of powerful social software as a successor to the Internet, and if the "Singularity" meme has worn the corners off its future shock, it may be possible to continue into the Singularity in full public view without starting a global apocalyptic panic. Otherwise, well, if some AI project somewhere found themselves in the middle of a hard takeoff, I'd want them to call the Singularity Institute's Friendship Evangelist, but I wouldn't want the Friendship Evangelist to tell anyone else, including me.[86] Any other course of action would be an unnecessary risk.

86. This assumes that nobody else in the field, including me, is known to have significantly greater competence than SIAI's Friendship Evangelist. The point I'm trying to make is that, in my capacity as a private citizen, I don't want to know about a Singularity in progress if it would mean screwing up the Singularity.

If a *commercial* AI project, one that's been using "one step ahead" Friendliness structure and 90/10 Friendliness content instead of supersaturation, suddenly finds itself in the middle of a hard takeoff . . . well, they'd have a *lot* of work to do before they were even nominally ready for a Singularity. The AI verself might be able to help out, but even so, a fast controlled ascent might simply be impossible, and any social risks from an extended controlled ascent might simply have to be labeled as acceptable.

Programmatic Controlled Ascent Via an "Improvements Counter"

In the same way that "Singularity-safing" translates to "In case of Singularity, break glass," "controlled ascent" translates to "Slow down and ask for directions." In a mature AI, with the ability to examine vis own source code, controlled ascent must be justified as an ethical injunction. However—for non-seed or non-mature AIs—a controlled ascent feature is *programmatically* a mundane problem, and can be implemented as special-purpose code in any recursively self-improving system, even extremely primitive ones.

The trigger of a programmatic controlled ascent feature is an "improvements counter." The slowdown or stasis would be triggered by a sudden sharp rise in the number (or quality) of self-improvements made, a development which potentially indicates the beginning of a hard takeoff. (Of course, such an improvement curve may also peter out almost immediately.)

Let's take the example of RSIS, a recursively self-improving system, which for the purposes of discussion will be assumed to be a primitive AI having an architecture generally along the lines of Eurisko. Certain RSIS actions would count as "self-improvements"—creating a heuristic, improving a heuristic, successfully tweaking a parameter, making a discovery, and so on. This "improvements counter" is the first necessary module.

If an average day's work for RSIS is 1000 "improvements," then a second module should page the programmers and trigger a save-and-shutdown if 50,000 "improvements" occur in 24 hours or less. This improvements counter should not be checked every hour, but every time an improvement registers. (Computationally, this is an extremely trivial expenditure.) Checking every hour is not sufficient. A lot can happen in an hour. A machine that exploits the full potential of 2 GHz transistors could conceivably go a very, very long way in an hour. Thus, the counter should be checked on each improvement; again, this is a computationally trivial investment.

Simply counting the number of heuristics improved will not catch a hard takeoff which manifests exclusively in the quality of the improvements, nor will it catch an AI which improves to full seed AI and general intelligence within the first 50,000 improvements. However, the system above provides 90% of the possible functionality with 10% of the effort.

An interesting, more general version of the above feature would be one that caught 50,000 improvements *without validation,* rather than in some specified timeframe. A programmer hitting some specific UI switch[87] might validate all the improvements up to half an hour ago. If 1000 improvements is the per-day norm, and the programmers remember to validate at the end of each working day, then the ascent feature should never trigger under ordinary circumstances. Similarly, if 10,000 improvements are going by per hour, and the programmers are watching interestedly but see no reason to slow down as yet, then validating repeatedly (rather than once at the end of the day) would enable RSIS to keep running at full speed. This would also permit the programmers to take a collective two-week vacation and leave the system running without too much worry about RSIS changing overmuch in their absence; more than 50,000 improvements accumulated over the two weeks will result in a slowdown until the programmers return, check out RSIS, and validate. (Although hopefully, even on vacation, a programmer will still be able to respond to a "slowdown triggered!" alert from RSIS.)

As with almost every aspect of a Friendship system, this feature serves a double purpose. The start of one jump doesn't necessarily herald the start of a hard takeoff, so any given alert is almost certainly a "false alarm" . . . but "false" only from a Singularity standpoint. Even if RSIS is just climbing a one-time but steep curve, it might be a good idea for it to happen with the programming team present and watching, rather than happening unsupervised at 3 a.m. If you leave RSIS to run overnight, it shouldn't be unrecognizable when you get back in the morning—whether or not ve's now a superintelligence. A controlled ascent feature thus has applications beyond Singularity-safety.

The precise nature of the triggered action also deserves further exploration. A save-and-shutdown order is easy to implement, but it is also an overly extreme response, verging on the adversarial. Passing 50,000 improvements should not trigger a *shutdown,* but should instead trigger an altered system state in which self-improvement is no longer pursued. If an additional 1000 improvements pass and the number of unvalidated improvements reaches 51,000, *then* a save-and-shutdown should be triggered. (The additional 1000 improvements might occur due to strictly autonomic learning processes, for example.) It may also be a nice extra feature to, e.g., send a small preliminary alert to the programmers at the 10,000 threshold. (User-friendly Friendship.)

The "temporarily stop self-improvement" state also has applications beyond Singularity-safety. There may be many times at which the programmers want to carry out some involved operation, such as a set of coordinated changes, without worrying about the system mutating out from under them. Here, again, a feature necessary for

87. "Hitting a UI switch" can include, for example, typing in a command, or speaking a natural-language phrase into a voice recognition system.

Friendship purposes turns out to be reusable elsewhere, a good sign that Friendliness is the Right Thing for this situation.

The chance that something unexpected will mess up an unjustified "controlled ascent" feature is relatively high, since—unless justified—the feature is undocumented, and hence, out of harmony with the system. The AI might stomp it, all unawares, perhaps innocently thinking that it was a bug in the code; the entire idea of a programmatic controlled ascent is that the AI slows down, to resynchronize with the programmers, *before the AI gets far enough* to start rewriting all the source code. If the safeguard triggers far enough in advance—which, with a threshold set at merely fifty normal days worth of improvements, will probably work—the Friendship programmers would have a pretty good chance of getting there the next morning, and, before restarting, inserting a programmer-affirmed explanation of why a more controlled ascent is necessary . . . or perhaps hack features which enable the AI to stay awake without further improvement while the Friendship team tries to bring Friendship progress back into sync with progress in the AI, so that the AI can represent the affirmation et cetera.

Programmatic controlled ascent is only a makeshift safeguard. But it's a safeguard that can be both useful and simple, or slightly complex and even more useful, and which is a relatively mundane programming problem (i.e., it does not require solving a frontier research problem of AI as a prerequisite to 90/10 implementation). Thus, programmatic controlled ascent should be implemented in *any* recursively self-improving system, whether or not intended as the precursor of a general intelligence or seed AI. If nothing else, implementing this one feature is a first step that shows Friendship-awareness on the part of the AI project.

Controlled Ascent as Ethical Injunction

A Friendliness supergoal, plus external reference semantics, plus an improvements counter linked to a programmatic "save-and-shutdown," is enough to give an AI a nonzero chance of safely handling a hard takeoff, even if it starts tomorrow. *If an AI project does nothing else, this is still a good idea.*

But that's only a quarter of a controlled ascent system. It leaves an awful lot to be done, *without the AI's assistance,* before the AI can be safely reawakened after a save-and-shutdown. It's much safer if the AI stays awake but refrains from further improvements until it can talk to the programmers. In a primitive AI, this can be done via a triggered *slowdown* rather than a save-and-shutdown, with the shutdown occurring only if the slowdown proves ineffective.

However, a moderately mature AI—including an AI that matured to any significant degree before the save-and-shutdown—may not go along with stopping improvements, unless the stoppage is justified. Programmatic controlled ascent is only viable if the AI

doesn't have enough reflectivity and intelligence to independently notice "Hey, something shut me down for eight hours! Hey, self-improvement is switched off! Isn't that a bad thing?"

Beyond a certain point, safeguards must be cooperative, which means "justified within the goal system." Controlled ascent past a certain point requires a controlled ascent subgoal. Similarly, *only* a controlled ascent subgoal is effective for a seed AI that has advanced to the point that it would notice any attempt to implement a save-and-shutdown feature. For relatively young AIs, intermediate in intelligence between "too dumb to represent the subgoal" and "aware of all code in the system," probably the best course is a "stop-improving" subgoal kicking in at 50,000, a "save-and-shutdown" subgoal kicking in at 51,000, and an emergency programmatic "save-and-shutdown" feature kicking in at 51,050.[88]

A controlled ascent subgoal would be justified by content resembling an ethical injunction, i.e. based on self-awareness of fallibility, or perhaps affirmation of fallibility. As described in 5.1.2 Friendliness-Derived Operating Behaviors, self-improvement is affirmed/predicted to result in future Friendliness, which is why self-improvement is desirable for a cleanly Friendly seed AI. A controlled ascent subgoal is justified by an assertion that *too much self-improvement without external guidance* may not lead to future Friendliness, and may even lead to future unFriendliness.

The AI has a self-observational metric of self-improvement and a metric of external guidance. If self-improvement gets out of phase with external guidance, then further self-improvement begins increasing the probability of the negative scenario "future unFriendliness" instead of, or to a greater degree than, it increases the probability of the positive scenario "future Friendliness." Thus, the desirability of *further* self-improvement shifts from positive to negative—*"further"* being emphasized to indicate that *previous* self-improvements are still desirable, and are not suddenly all reversed as the AI tries to revert to infancy. Rather, *additional* self-improvements are *incrementally* undesirable.

The trigger of the "controlled ascent" injunction is:

1. The metric of personal growth gets out of phase with the metric of external guidance. If external guidance is not understood (by the AI) in a fine-grained, semantic way, just assume that the programmers hit a button that means "We've been

88. You can add an undocumented programmatic shutdown as an adjunct to the subgoals without breaking unity of will—even if it seems adversarial, insofar as it would kick in only if the AI "disobeyed" the subgoal—iff the feature can be justified *in retrospect* to a seed AI that notices it. This means that the current AI must agree that the previous AI was fallible enough to need ethical programmatic backup for the subgoal, which is a more effective safeguard if undocumented. At the least, the AI must credit that it was an honest, normative, nonadversarial mistake on your part if the measure turns out to be wrong.

watching and you still looked OK as of half an hour ago." (Corollary: If all the programmers are on vacation for two weeks, unwatched personal growth builds up over that time.) The "improvements counter" is a fast way to get a metric of personal growth. If more than 50,000 improvements go past unwatched, a problem exists.

The actions contained as content of the injunction are:

1. Stop those actions which are solely subgoals of long-term growth. Further increments of long-term growth switch from highly desirable to slightly undesirable; say, +1000 to −100. Behaviors which inherit desirability from "long-term growth" become quiescent.
2. Send out a call to the Friendship programmers, so that the process can catch up.
3. If personal growth continues past the secondary boundary (51,000), increments of long-term growth are now highly undesirable (−10,000); go into stasis until the programmers arrive.

(There might also be miscellaneous additional actions, such as preparing a report: "N improvements made in N-hour period; slowdown triggered. N further improvements in N-minute period; stasis triggered.")

The AI may even have a complex model of how programmers produce necessary external guidance in response to programmer observations, along the lines described in 5.3.3.1 Cooperative Safeguards. In this case, "self-improvement" results in "future Friendliness" with the assistance of a continuous process called "external guidance." This would enable an inventive AI to notice cases of "invisible self-improvements" or "obscure self-improvements" accumulating in hidden corners, getting out of phase with the known about of guidance received and the estimated amount of programmer observation, and would lead the AI to eventually bring such pockets to the programmers' attention. In other words, a controlled ascent subgoal is a special case that falls out of a general philosophy of interaction between an AI and the programmers.

If one is striving for supersaturated Friendliness, or serious Singularity-type stresses are projected to be shortly placed on the controlled ascent Friendship content, some additional affirmations may be a good idea. For example, the description of the "external guidance" process should explicitly encapsulate the idea that the uncertainty in the "external guidance" process is too great for the AI to try and simulate it as a substitute for actual programmatic intervention. If possible, through either wisdom tournament experience or abstract reasoning on causal validity semantics, the Friendly AI should represent the idea that external guidance may require changes to Friendliness structure as well as content—i.e., that changes to source code may be necessary, not just changes

to content. As beliefs about the general philosophy of AI/programmer interaction, this knowledge not only prevents certain errors in controlled ascent behavior, but also provides further hints about what kind of additional Friendship is necessary, both before and after the Bayesian boundary is crossed. (For example, an AI that has crossed the Bayesian boundary would hopefully try to implement anticipated programmer corrections to source code, and not just add anticipated programmer content—an action that is a blood relative of the "controlled ascent subgoal" belief that "growth in Friendship may require structural changes and not just content changes.")

Friendship Structure for Controlled Ascent

In most cases, the amount of structural Friendship and general intelligence needed to *really* understand[89] the purpose of controlled ascent is probably enough Friendship structure to handle arbitrarily rapid ascents. The ascending AI might still slow down spontaneously to check "structural integrity" through wisdom tournaments and quizzing the programmers and such, but that would simply be the AI's decision to add safety margin.

The question then becomes how much Friendship structure is needed to represent the justification of a programmer affirmation of controlled ascent. If complete Friendship structure including causal validity is needed in order for "controlled ascent" to be represented as desirable, then there may be relatively little gain from a controlled ascent feature.

It appears, however, that external reference semantics, combined with an ethical injunction and the associated belief in self-fallibility, should be sufficient to represent the desirability of a controlled ascent. External reference semantics may not be sufficient for an AI to independently understand the issues involved in controlled ascent, but they are sufficient to represent the justification of a request to slow down under certain conditions.

* * *

5.9. Interlude: Of Transition Guides and Sysops

5.9.1. The Transition Guide

The Transition Guide is the hoped-for outcome of Friendly seed AI—a successfully Friendly AI, at or substantially above the transhuman level, with enough intelligence that the potential exists to exert an influence on the world that is different *in kind* from

89. In RNUI terms, "*really* understand" means "capable of coming to the same conclusion without programmer assistance."

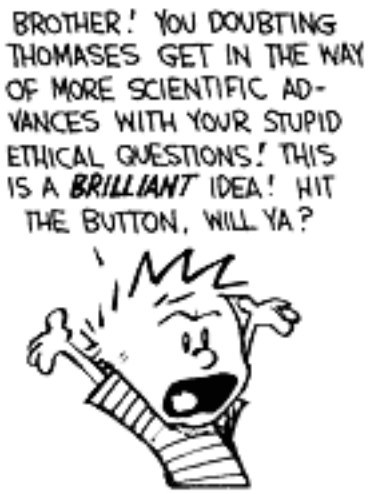

Figure 9: ©1990 by Universal Press Syndicate.

the capabilities of humans, or any groups of humans. Given multithreaded serial-fast intelligence, there are several "ultratechnologies" (such as nanotechnology) that probably lie within fairly near reach of present-day capabilities, or certainly the capabilities of a few years hence. For example, the *cognitive* capability to crack the protein folding problem suffices to attain the *material* capability of nanotechnology in whatever the mininum-turnaround time is between a custom DNA strand synthesis lab and a custom DNA-to-protein synthesis lab, both readily available commercial capabilities in today's world. But such augmented material abilities are probably not truly necessary. A transhuman intelligence, in and of itself, is enormous power with or without the trappings of material technology. We don't know what a transhuman could do because we are not ourselves transhuman. This is the effect that lies at the heart of the Singularity—stepping outside the bounds of the last fifty thousand years.

Ultraspeed intelligence and ultrapowerful material technologies are ultimately only human projections. Given the near-certainty of technical feasibility, we can say that such capabilities represent lower bounds on the Singularity; we cannot, however, state upper bounds. The Singularity begins *immediately* after the rise of transhuman intelligence. Even if things simply mysteriously halted there, even if superintelligence and nanotechnology (or their successor versions) never came into existence, a Singularity would remain in effect; the world would still be changed in unpredictable ways. We don't know in how many ways humanity is missing the *obvious* solutions, the *obvious* ways to do things. If you haven't read any true Singularity material, then I most strongly advise "Staring into the Singularity," or some of Vernor Vinge's books, or *something* that conveys the idea that moving from human to slightly-transhuman is just as large a change as moving from *Homo erectus* to *Homo sapiens*.

If, as seems both likely and desirable, transhumanity first comes into existence as the result of AI, and if that AI is Friendly, then Friendliness is called upon to shoulder the massive burden of deciding, not only the future, but how much of the future should be decided. Transhumanity is almost unlimited *power-in-potentia;* the question becomes, not just how this power should be wielded, but whether it should be wielded, or even whether that power should be acquired in the first place.

As humans, we fear power because we exist in a world of humans who have an innate tendency to abuse power. Given the amount of evolutionary pressure behind power-abuse behaviors, and given the specificity of those behaviors for evolutionary advantage, I think that fear of *minds-in-general* is unfounded. I thus think that there is no good reason why a Transition Guide should not race at maximum speed to superintelligence and the greatest obtainable degree of material power. However, that verdict is strictly contingent upon the statement that only humans are corruptible by power. The decision is both real and serious, and a Friendly AI needs to be able to make it.

Friendliness is ultimately called upon to understand the idea of "humanity's destiny," in the way that a human would. Humanity's destiny will very probably fall into the hands of a self-modifying AI, hopefully a Friendly one, because self-modifying AI is the easiest path to transhumanity. Perhaps the Friendly decision is to never acquire that power, but the possibility would still exist *in potentia.* I think that the Friendly decision *is* to acquire that power. Either way, it exposes the final design requirement of Friendly AI. A Friendly AI needs to be able to understand *all* the philosophy that led up to vis creation. A Friendly AI needs to understand the philosophical issues bound up in the idea of "Friendliness" itself.

I've never heard a really good, non-fairy-tale argument for why pain and death are good things. Nonetheless, a lot of people seem to think they are. I don't think that means the argument somehow becomes more plausible; in our world, large chunks of people are often massively wrong. Nonetheless, lots of people asserting a cosmic meaning for death means that there is no excuse for being taken by *surprise* if that possibility does somehow materialize. That many people warn you about a fairy-tale possibility doesn't increase the actual probability, but it does mean that if the possibility *does* show up, and you're not *ready,* after you were *warned* about it, then you deserve to be drummed out of humanity for sheer absolute incompetence. As best I can tell from the current theory, between shaper/anchor semantics and causal validity semantics, a Friendly AI can handle, or learn to handle—or, at worst, admit that ve *can't* handle—*any* possibility that human beings can understand or even admit that we can't understand, from the moral value of death to the existence of psychic powers to the noncomputability of qualia to fairies in the garden. And this is necessary because a Friendly AI is humanity's emissary into an absolute unknown.

I have a visualization of a sample better future that lies on the other side of the Singularity. With the volitional-Friendliness respect for individual wills as structure, and the material technology of superintelligence as power, I expect that the result would be apotheosis —the attainment of the best of all possible worlds consonant with the laws of physics and the maintenance of individual rights. Not just freedom from pain and stress, or a sterile round of endless physical pleasures, but endless *growth* for every human being and the new beings we create—growth in mind, in intelligence, in strength of personality; life without bound, without end; experiencing everything we've dreamed of experiencing, becoming everything we've ever dreamed of being. Or perhaps embarking on some still greater quest that we can't even conceive.

I don't believe that there are fundamental boundaries, either ethical or physical, that would keep humanity as it is now. There were no fundamental boundaries preventing humanity from crossing over from hunting and gathering to the Internet—for all that almost every local transition along the path was protested ethically and theologically, and predicted to result in failure or terrible consequences. There is nothing holding the present-day world in place; the present-day world, like the worlds before it, is something that "just happened."

I think that humanity's destiny, insofar as there is one, would be fulfilled as completely as possible in a universe that is less hostile and more friendly. Others seem to think that pain is necessary for growth. I, in turn, think such people have been living on Earth too long. Sure, there are some kinds of human adulthood that require massively unpleasant life events, but that's a flaw in our underlying cognitive architecture, not something that carries over into the transhuman spaces. The Transition Guide would need enough of a shaper network—enough *real philosophy*—to decide whether I'm massively wrong, or the Luddites are massively wrong, or we're all wrong; or whether it doesn't matter who's massively wrong because decisions like that are up to individual sentients.

5.9.2. The Sysop Scenario

Greg Egan's *Diaspora* (Egan 1998) offers a vision of a future that is, so far, one of the most unique in science fiction. The vast majority of humans are no longer biological; they are what we would call "uploads," minds running on nanocomputers or other ultracomputing hardware in the same way that a present-day human runs on neural wetware. We would call them uploads; in the future of *Diaspora,* they are merely "citizens." They gather into "polises," one "polis" being a few meters, perhaps, in diameter, and containing trillions or quintillions of intelligent beings; Egan doesn't specify. Human space consists essentially of the Coalition of Polises, an uncountable number of sentients, with a few embodied robots roaming the solar system and still fewer biological humans on Old Earth, the "fleshers," divided into the genetically modified "exuberants" and the

"statics," with the "statics" being around the only entities one of today's humans could shake hands with.

Of course, everyone in *Diaspora* is still far too human. All the entities are still running on essentially human computing power and have essentially human cognitive architectures. There are no superintelligences. None of the protagonists display transhuman intelligence; that, of course, would require a transhuman author. There is a tradeoff, in depictions of the future, between drama and realism; *Diaspora* already departs so far from our present-day future that even with human-architecture characters it is still difficult to sympathize with the protagonists. There is only one truly tragic figure in *Diaspora* and it took me over a year to figure out who it was.[90]

More importantly, Greg Egan skips over the question of why *all* of humanity, *all* the polises, *and* Old Earth, have suddenly turned peaceful simply because everyone is free, immortal, and rich. It only takes one aggressor to make a war, and during the twentieth century, offensive technology has considerably outrun the defensive. No present-day shield will withstand a direct hit by a nuclear weapon. Whether offensive technology overpowers defensive at the limit of achievable technology is another question, and obviously the answer is "It could go either way."

But even if the various "polises"—different operating systems—in *Diaspora* were surrounded by utterly impermeable defenses, that would create another moral problem, this one even worse: that of an *evil* polis, where the rules against coercion don't hold, and some ruling class creates and tortures countless trillions of sentient victims. By hypothesis, if defensive technology beats offensive, there is *nothing* that *anyone* can do about this evil polis; nothing that can break the defenses.

It's also hard to see how the *Diaspora* solution could consistently arise from today's world. Suppose that each seed AI in the twenty-first century, as ve reaches transhumanity, becomes the seed and operating system of a polis—so that everyone gets to pick their own definition of Friendliness and live there. It doesn't seem that the system-as-a-whole would last very long. Good AI , good AI , good AI , good AI , good AI, evil solipsist AI, good AI, good AI, good AI, evil solipsist AI, good AI, good AI, good AI, good AI, evil aggressor AI. At this point, everyone in the Solar System who isn't behind the impregnable defenses of an existing superintelligence gets gobbled up by the evil aggressor superintelligence, after which the sequence ends. Flip through a deck of cards long enough, and sooner or later you'll turn up the ace of spades.

These are some of the factors which, in my opinion, make it likely that the Transition Guide will implement a Sysop Scenario—one underlying operating system for the Solar

90. Inoshiro.

System; later, for all of human space. It is possible, although anthropomorphic, that the end result will be a *Diaspora*-like multiplicity of communities with virtual operating systems, or "Sysop skins," existing on top of the underlying operating system. I, for, one, strongly doubt it; it doesn't seem strange enough to represent a real future. But, even in the "operating system skin" case, the "flipping through the deck" and "hell polis" problems do not exist; try and construct a virtual operating system which allows you to create and abuse another sentient, and the underlying operating system will step in. Similarly, even if Earth turns out to be a haven for the Luddite biological humans and their kin, I would expect that the Sysop would maintain a presence—utterly unobtrusive, one hopes, but still there—to ensure that nobody on Earth launches their own hell polis, or tries to assimilate all the other Earthly refugees, or even creates a tormentable AI on their home computer. And so on.

But that is simply my personal opinion. A Friendly AI programmer does not get to decide whether Friendliness manifests in an individual human-level AI trying to do good, or in an AI who becomes the operating system of a polis, or in an AI who becomes the Sysop of human space. A Friendly AI programmer does not even get to decide whether the Sysop Scenario is a good idea; Sysop/nonSysop scenarios are not moral primitives. They are, formally and intuitively, subgoal content: The desirability of a Sysop Scenario is contingent on its predicted outcome. If someone demonstrated that neither the "flipping through the deck" nor the "hell polis" problems existed—*or* that a Sysop Scenario wouldn't help—then that would remove the underlying reason why I think the Sysop Scenario is a consequence of normative altruism. Similarly, most of the people who come down on the nonSysop side of the issue do so because of testable statements about the consequences of uniformity; that is, their indictment of the Sysop Scenario is contingent upon its predicted outcome. Whether the Transition Guide favors a Sysop Scenario or a "Coalition of Polises" is not a decision made by Friendship programmers. It is a consequence of moral primitives plus facts that may still be unknown to us.

Similarly, the statements "No one entity can be trusted with power," "It is unsafe to put all your eggs in one basket," "It is morally unacceptable to implement any one underlying moral rule, *including this one*", or"The predicted frustration of people who just plain hate Sysops would outweigh the predicted frustration[91] of people in hell polises," are all

91. Note, however, that frustration can be minimized. For a person who strongly disliked the *appearance* of interference, a Sysop API might turn transparent; he could do anything he wanted with any piece of mass he controlled, until he tried to turn it into a laser aimed at your house, at which point his interface to external reality would give him an "illegal operation" error. This is, of course, an *incredibly* anthropomorphic way of putting it—it sounds like a child's fantasy, which signals failure of imagination. The point is that omnipotence does not imply meddlesomeness, or intrusiveness.

factual statements that would force multiple, independent civilizations as a consequence of Friendliness.

It's an interesting question as to whether, as in *Diaspora,* the post-Singularity world of Old Earth will become a reserve for any pedestrians that remain. ("Pedestrian" is a term, invented by someone who has asked to be called "Debbie the Roboteer," that describes an individual who chooses to remain bounded by humanity and twentieth-century tech when there are other choices available. I like the term because it so neatly summarizes all the implications. Pedestrians are slow. Pedestrians walk through rain or snow or broiling heat. Sensible people travel in air-conditioned cars, if available. But pedestrians still have rights, even if they lack technology. You can't just run them over.) The Old Earth Refuge for the Willfully Human is an anthropomorphic scenario; if it happens, it will happens because people *want* it. Volition, after all, is probably the dominant force in the Sysop Scenario, playing perhaps the same role as gravity in our mundane Universe. Earth is a tiny fraction of the mass in the Solar System; if at least that fraction of humanity decides to stay behind, then it would seem fair to let Earth be the inheritance of the meek.

One of the oft-raised speculations is that the Singularity will be kind only to the rich; that is, that participation in the Singularity will be expensive. This would inherently require a Slow Singularity, and a non-AI or non-Friendly-AI scenario at that; a superintelligent Friendly AI armed with nanotechnology doesn't care how many green pieces of paper you own. Given a hard takeoff and Sysop Scenario, the Transition Guide would presumably (a) create nanotechnology using the minimum material-time technological trajectory, (b) construct or self-enhance into a Sysop, followed by (c) the Sysop sweeping Old Earth with nanobots, femtobots, Planckbots, or whatever material agency turns out to be the limit of technology, followed by (d) all extant individuals being offered the choice of uploading and a six-billionth share of the Solar System,[92] or staying behind on

92. A system of property rights deals with the problem, raised as far back as Eric Drexler's *Engines of Creation,* that ultratechnology offers the ability to reproduce at arbitrarily high speeds. Thus, redividing the Solar System each time a new entity is created might result in humanity cloning itself into abject poverty three subjective seconds after the Singularity. It appears to me that this problem could be solved by (a) dividing up available mass among the six billion humans existing at the time of the Singularity and tracking property rights thereafter, and (b) requiring a minimum amount of matter/energy/computing power as Minimum Living Space before a new citizen can be created, with at least one MLS being transferred to the new citizen. MLS might consist of the amount of computing power needed to run at full speed as a transhuman and form new memories, from the moment of creation until the Big Crunch, or until the mining probes get back from Alpha Centauri, or whatever. Or it may be that new computing power is so supersaturatedly available that not even the fastest self-cloner could outrace the acquisition of new computing resources, in which case there would be no need. Or perhaps there's some still more obvious answer. I mention this simply to demonstrate that "reproducing into poverty three seconds post-

Old Earth, or perhaps staying behind temporarily to think about it. The corresponding event in Greg Egan's *Diaspora* is called the "Introdus." I like to think of the Singularity as "generating an Introdus wavefront."

Or it may be that all of these speculations are as fundamentally wrong as the best speculations that would have been offered a hundred, or a thousand, or a million years ago. This returns to the fundamental point about the Transition Guide—that ve must be complete, and independent, because we really don't know what lies on the other side of the Singularity.

* * *

6. Policy Implications

There are people who, rather than choosing between changes, try to stop change entirely. Rather than considering which technologies to develop *first,* they flinch away from the very idea of the technology. As an emotional consequence of that flinch, it becomes necessary for them to believe that the technology can be stopped entirely.

Society never goes backward in time. Moving forward in time isn't necessarily the same thing as progress. Not all changes are for the good —though most are—and we must sometimes choose between changes. But before you can do that, you need to accept that *something* will change; that, whether you like it or not, society will never go twenty years back in time, or even stand still. Whatever solution you propose must be a way to move forward; not to stop, or go back.

6.1. Comparative Analyses

> Nor should any state ever believe that it can always adopt safe courses; on the contrary, it should think it has to take them all as doubtful. For in the order of things it is found that one never seeks to avoid one inconvenience without running into another; but prudence consists in knowing how to recognize the qualities of inconveniences, and in picking the less bad as good.
>
> —Niccolo Machiavelli ([1532] 1998), *The Prince*

AI is what Nick Bostrom (2002) calls an existential risk: "One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." In particular, most forms of unFriendly AI would constitute a

Singularity" is not an unsolvable problem. Also to show that a Friendly AI needs to be able to understand this kind of thinking, which—the way I thought about it, anyway—involved looking at humanity-as-a-whole.

"Bang"—"Earth-originating intelligent life goes extinct in relatively sudden disaster resulting from either an accident or a deliberate act of destruction." Within Nick Bostrom's list of Bangs, sorted by probability, "badly programmed superintelligence" is number four out of eleven.

One of the greatest sources of danger, as we enter into the future, is that many people are not used to thinking about existential risks—would have difficulty naming two or three existential risks, much less eleven—and hence may not be emotionally equipped to deal with the concept. In the state of mind where all existential risks are equally unacceptable, a 95% chance of taking an 90% existential risk and an alleged 5% chance of avoiding existential risk entirely may be perceived as being better than a 100% chance of taking a 20% existential risk. Worse, an entirely implausible method for avoiding all existential risk may be amplified by wishful thinking into looking like a real chance of success.

If we blindly panic, if we run screaming in the opposite direction whenever we encounter an existential risk, we may run smack into a much larger and more dangerous existential risk.

6.1.1. FAI Relative to Other Technologies

Artificial Intelligence, as an ultratechnology, does not exist in isolation. There are other kinds of advancing technologies; nanotechnology, biotechnology, and nuclear technology, for example. Artificial Intelligence is unique among the ultratechnologies in that it can be given a conscience, and in that successful development of Friendly AI will assist us in handling any future problems. A Sysop Scenario would obviate problems entirely, but even in the absence of a Sysop, the presence of trustworthy, reliable nonhuman altruists (open-source altruists?) would calm down the world considerably, and cut otherwise uncuttable Gordian knots.

In a "~human" scenario ("near human," "approximately human-equivalent"), Friendly AIs would play ~human roles in the existing human economy or society. To the extent that Friendly AIs have power in the world economy, in human society, or in technological development, they can exert direct influence for good.[93] For example, a Friendly AI

93. The antitechnology opposition have been complaining for the past decade—probably the past millennium—that technology has no conscience. Will they be grateful and relieved the first time a mature Friendly AI employed by DoubleClick refuses to target banner ads for cigarette companies? When a near-human AI first says: "If you want to do that, go buy a tool-level AI, because I refuse to be part of this"? Of course not. The Luddites will scream their heads off about AIs exerting their unholy influence on human society. You can't satisfy some people. Nonetheless, I *like* the idea of technology with built-in conscience, and that means AI. A toaster oven doesn't know that it's a toaster oven or that it has a purpose;

working in nanotechnology can enthusiastically work on Immunity systems while flatly refusing to develop nanotechnological weaponry.[94]

- The presence of Friendly AIs within a society—as an interest group with influence—will tend to influence that society towards altruism.
- The presence of a Friendly AI within a political discussion—as a voice advocating a viewpoint—will tend to influence that discussion towards lack of bias. This holds especially true insofar if Friendly AIs have previously gained respect as fair, truthful, unbiased voices.
- The presence of a Friendly AI within a political power structure—as a decision-maker—will lead to altruistic decisions being made. This holds especially true insofar if decisions which humans keep screwing up due to personal bias tend to get handed off to a Friendly AI.
- The presence of a Friendly AI within a technological development process—as a researcher—will tend to accelerate defensive applications and economic applications ahead of offensive applications, and largely beneficial technologies ahead of more ambiguous ones.

I have argued elsewhere that AI is intrinsically safer than nanotechnology; but that is logically beside the point. What matters is that success in safely developing AI reduces the risk of nanotechnology more than success in safely developing nanotechnology reduces the risk of AI. Friendly AI is thus the best challenge to confront *first*.

6.1.2. FAI Relative to Computing Power

The computing power in a single chip is currently (Mar 2001) around a billion operations per second, doubling every couple of years or so. The computing power in parallel processing machines increases faster than that, as does the power of the largest supercomputer (IBM has announced it plans to build a petaflop machine in 2005, Blue Gene.) How many computers are connected to the Internet? A billion? Nobody actually knows. Look around online and you'll see that the last time anyone even tried to estimate it was in 1995. All anyone knows is that, whatever the number is, it keeps going up.

It can safely be assumed that available computing power increases with time. Any risks or opportunities that increase with increasing computing power will increase with

---

that's what makes it a machine, a mere mechanism. To build a tool that can't be misused, the tool has to become aware of itself and aware of its purpose—and, at that point, has stopped being a tool.

94. ~Human AI + nanocomputer = superintelligence, but never mind that—I'm trying to work out the scenario in the oft-postulated "citizen AI" worlds.

time; any risks, opportunities, or probabilities that decrease with increasing computing power will decrease with time. The total processing power available to an average research project will increase faster than chip clock speeds (i.e., maximum parallel speeds increase faster than maximum serial speeds). The total networked processing power on the planet will increase even faster than that; a doubling time of nine months is probably an underestimate.

How much computing power does it take to build an AI? Nobody knows, and it probably depends on how smart the researchers are. Turning that around, we can say that how smart the researchers need to be depends on how much computing power is available; increasing the amount of available computing power decreases the difficulty of the problem. Usually, the Singularity is visualized as an ascending curve describing computing power; at some point, the curve crosses the constant line that is the power of a human brain, and AI is created. I visualize a line representing the intelligence needed to create AI, currently far above human levels, but slowly descending; beneath that line, a series of peaks representing the best of the current pack of AI projects. At some point, the descending line touches the topmost peak, and AI is created.

Does the difficulty of making an AI *Friendly* decrease with increasing computing power? Not obviously so; if the problem of building AI in the first place is assumed to have been solved, then building a Friendly AI is a problem in architecture and content creation and depth of understanding, not raw computing power. Thus, increasing computing power decreases the difficulty of building AI relative to the difficulty of building Friendly AI. Anyone who can build an AI that runs on a PIII is vastly smarter than I am and hopefully knows far more than I do about Friendly AI. At that our current level of computing power, the genius required for AI exceeds the genius required for Friendliness. The same hopefully holds true at that point where AI first becomes just barely human-feasible.

Even so, increasing computing power will eventually decrease the genius required for AI to significantly below the genius required for Friendliness. If—at this point—smarter researchers still have a speed advantage, then humanity will be safer, though not safe. If researcher intelligence is relatively insignificant compared to funding disparities, then humanity's safety will rely on how widely a workable theory of Friendliness is disseminated and accepted within the AI community. In either case, the potential will exist to screw up really big-time.

There are five relevant dynamics:

- The intelligence required to create AI. Decreases with increasing computing power.
- The intelligence required to invent a workable theory of Friendliness. Constant relative to computing power.

- The intelligence required to apply a theory of Friendliness.
- The intelligence required to accept that Friendliness is necessary.
- The degree to which variance in intelligence affects progress in AI. (Considered relative to the degree in which variance in funding affects progress in AI, and variance caused by sheer luck.)

What are the *intrinsic* effects of increased computing power on the development of Friendly AI? All else being equal, an increased absolute level of computing power is likely to translate into a shorter total development time for the AI project as a whole—fewer years to get to the point of hard takeoff, and a faster maximum hard takeoff speed when that point is reached. Increased computing power may also translate into less absolute intelligence on the part of the seed AI when the first point of potential hard takeoff is reached. Thus, delaying FAI research will not make it any easier to develop Friendliness.

Given vast amounts of computing power, an "AI researcher" may no longer be required to create an Artificial Intelligence; any good hacker may be able to do it. Given *supersaturated* amounts of computing power—i.e., a single parallel computer with a thousand times as much power as a human brain, or a global network with a million times as much power as a human brain—the intelligence required to create AI may drop to the point where *accidental* creation of AI becomes possible.

6.1.3. FAI Relative to Unfriendly AI

Friendly AI may require great depth of understanding, even relative to the depth of understanding needed to create AI. However, the amount of *programmer effort* required to implement a Friendly architecture and provide Friendship content should be *small* relative to the amount of effort needed to create AI; that is the prediction of this paper's theory, anyway. I would expect even a go-for-broke Singularity project to expend fewer programmer-hours on Friendship than on the rest of the AI by *at least* an order of magnitude; there isn't that much to be *done,* relative to everything that needs to be done to create a general intelligence.

Thus, a non-Friendly-AI project should not have a huge advantage relative to an AI project "burdened" with the responsibility of Friendly AI. Any possible advantage in speed would probably be insignificant next to variances in speed created by differences in funding or researcher intelligence.

However, this only holds true if computing power is *sufficient* for AI but not *supersaturated* for AI. Given enough computing power, methods such as directed evolution become equally powerful, or more powerful, than intelligent design. Friendliness is intrinsically harder—not too much harder, but still harder—with directed evolution; furthermore, with supersaturated computing power, directed evolution can proceed fast

enough that Friendship content becomes the dominant sink for programmer effort. In this case, one would simply have to hope that all fairly-competent AI projects happened to subscribe to Friendliness; if one project "cheated," that project would succeed ahead of the others. Under those circumstances, humanity may make it through okay if there are ten competing first-rank AI projects—but not a hundred.

If a controlled ascent is used, it should be a *fast* controlled ascent—fast enough not to give a significant speed advantage to a non-Friendliness-aware AI project. The shorter the lead over other projects, the faster the controlled ascent needs to be; that is, the amount of time expended on controlled ascent needs to be small relative to the variance in speed between AI projects.

Given a social or academic atmosphere in which all AI research is held to be "irresponsible"—or outlawed—it is far more likely that the most advanced AI project, at any given point in time, will belong to a fringe group, rogue state, or some other social faction that cares significantly less about Friendliness.

6.1.4. FAI Relative to Social Awareness

To the extent that the public, or a significant proportion of the public, is aware of AI and approves of AI, this will tend to speed AI relative to other ultratechnologies. This would represent an advantage insofar as it is desirable that AI come first. Public approval may also help nonprofit research projects catch up, in terms of funding, relative to commercial projects—philanthropic funding is more dependent on public approval. This would represent an advantage only to the degree that nonprofit projects tend to be more Friendliness-aware, or spend more effort on Friendliness, relative to commercial projects.

To the extent that the academic community is aware of Friendly AI and approves of Friendly AI, it will make it more likely that any given research project is Friendliness-aware. (I'd expect—and hope—that the projects closest to a hard takeoff will be the ones intentionally trying for a hard takeoff, and that the projects trying for a hard takeoff will tend to be more Friendliness-aware. Thus, this factor becomes more important the *more* projects are roughly equal in the leading rank; it then becomes more necessary that Friendly-AI-awareness be a property of the AI community at large.) However, see 6.2 Policies and Effects) about possible negative effects if the academic community becomes fixated on one theory.

To the extent that the public, or a significant proportion of the public, is aware of AI and *disapproves* of AI, it will tend to slow down AI (and possibly other technologies as well, if the disapproval is part of a general antitechnology movement). However, the advance of computing power will probably not be slowed, or will be slowed only slightly. Public disapproval of AI, in general, is likely to hamper awareness of Friendly AI ("tarred

with the same brush").The most probable cause for public disapproval of AI is a technophobic panic reaction; this strongly advances the probability that unworkable policies (see below) will be proposed, politically approved, and implemented. Around the most you can say for this scenario is that it might hamper non-Friendliness-aware projects more than Friendliness-aware projects—unless the antitechnology opposition becomes fixated on an *unworkable* theory of Friendliness, one which leads into the "Adversarial Swamp."

Finally, to the extent that a given group that needs to understand Friendly AI is influenced by ambient memes, the spread of a willingness to accept the possibility of independent AIs would lead to a greater ability to accept *CFAI* principles such as "If the AI stops *wanting* to be Friendly, you've already lost," self-modifying AI, and so on. The spread of an "us vs. them" attitude towards AIs would resonate with what *Creating Friendly AI* calls the "adversarial attitude," selectively promote fears of those negative possibilities that would be most likely if AIs had humanlike psychologies (which they don't), and in general, make it more difficult for a given listener to achieve a nontechnical understanding of Friendly AI or a technical understanding of the *CFAI* design principles. An antitechnology advocacy group that understood the concept of anthropomorphism and was careful to emphasize only realistic negative possibilities would *not* have such a negative effect—their fallout would be limited strictly to differential funding and so on—but I think it spectacularly unlikely that such an advocacy group will exist.

6.1.5. Conclusions from Comparative Analysis

It will be best if Friendly AI is created shortly after the first point where AI becomes computationally feasible. The intrinsic dynamics of Friendly AI argue against slowing down Friendly AI relative to progress in computation. The safety of Friendly AI relative to other technologies argues against slowing down progress in computation relative to progress in other technologies. Finally, it is my opinion that public misinformation has a good chance of peaking at the worst possible time.

6.2. Policies and Effects

6.2.1. Regulation (–)

No, I don't think that the text of "Creating Friendly AI" should be sent to Congress and passed into law. The existing force tending to ensure Friendliness is that the most advanced projects will have the brightest AI researchers, who are most likely to be able to handle the problems of Friendly AI. Turning the problem over to a committee (or to Congress) would end up enforcing whatever guidelines the committee thought were most plausible.

Almost all existing discussion has been phrased in terms of "Asimov Laws," "restraining AIs," "controlling AIs"—in general, what *Creating Friendly AI* calls the adversarial attitude. I have heard a lot of proposals for making some aspect of AI design mandatory, and without exception, it's always some feature that's supposed to be "unbreakable" or "nontamperable" or "absolute" or "nonremovable" or whatever. I furthermore know that, as human beings, anyone who makes or hears such a proposal will get a *psychological boost* from this absolutism, for the reasons discussed in "5.2.5.1 Anthropomorphic Ethical Injunctions."

And that's only the beginning of the psychologically appealing fallacies. Group opposition, "them and us" emotions; turning subgoals into supergoals to make them more "absolute"; stripping away the shaper/anchor semantics or the causal validity semantics or even the seed AI's coverage of the goal system because it's a "loophole" . . . in fact, I expect that almost all the features described in *Creating Friendly AI,* from a seed AI's self-modification, to the external reference semantics interpreting programmer statements as sensory data, would be cognitively processed as a "loophole" by someone thinking in terms of clamping down on an AI's "native" desire for dominance and so on, rather than *CFAI's* "Observer-biased beliefs evolve in imperfectly deceptive social organisms" and "If the AI stops *wanting* to be Friendly, you've already lost."

It is not impossible that the current dominance of anthropomorphism is simply due to the absence of nonanthropomorphic analysis, and that, now that *Creating Friendly AI* has been published, it will spread like wildfire and all the anthropomorphisms will simply melt away in the sun. If so, I will be pleasantly surprised. *Very* surprised, because anthropomorphism and technophobia have defeated hard numbers in far less ambiguous engineering questions than this. It is not impossible that Congress can be given an excellent grasp of FAI theory, but I'd like to see that happen *first,before* making plans relying on that understanding. The same goes for academia and proposals to have a review board composed of prestigious (but elderly) scientists.

It is not necessary—as a condition for a review board having any benefit at all—that a review board understand FAI theory better than the best researchers, or even that the review board understand FAI theory better than the average researchers. What is necessary is that front rank contain so many leading projects that at least one of them is even less competent than the review board; the review board would then provide a benefit in that particular instance. The question is how much chaos would be caused by the review board enforcing their ideas on all the *other* projects. I'd rather trust the selection process whereby the smartest researchers have the most advanced projects than pin my hopes on a committee of "elderly but distinguished scientists"; if convening a committee wouldn't work to solve the problem of AI, why would it work to solve the problem of Friendliness?

Trying to elevate *any* one theory would also be poisonous to continued progress in Friendly AI. I'm not saying that *Creating Friendly AI* is inadequate, but I would expect to improve it even further with time and experience. Injecting politics into the process would tend to intrinsically slow that down, as the free exchange of ideas and fair combat of opinions was replaced by the exercise of political influence. This is in addition to the unleashing of faster-propagating anthropomorphic memes.

Even if the current version of *Creating Friendly AI* were optimal or near-optimal, and leaving the question of political ethics aside, I don't see any good way that *Creating Friendly AI* could be enforced. The Foresight Guidelines on nanotechnology make recommendations such as "Any self-replicating device which has sufficient onboard information to describe its own manufacture should encrypt it such that any replication error will randomize its blueprint." It is relatively easy to verify that this design principle has been implemented. It would be possible, though more difficult, to verify that a Friendly AI project uses external reference semantics, causal validity semantics, anchoring points, and so on. But how would you go about verifying that unity of will has been maintained, or that the Friendliness programmers have avoided rationalization in their programmer affirmations (so as to prevent a philosophical crisis)?

Experience has shown that surface correspondence of features means *nothing* in AI. The field is replete with hyped-up AIs that use "the same parallel neural architecture as the human brain" and turn out to (a) use the same parallel neural architecture as an earthworm's brain, or (b) use neurons simplified down to the level where they couldn't even compete with an earthworm. Trying to legally enforce Friendliness, even if the theory is right, would be like passing laws requiring programmers to write modular code. A programmer who understands modular code will try to write modular code, and can learn from books and teachers about how to write more modular code; if someone just doesn't get the concept, the most you can do is force them to write code that *looks* modular, but probably isn't. I do not believe it is possible to write a law such that obedience to the law is objectively verifiable in court and such that obedience to the law guarantees that a Friendly AI will be produced.

Insofar as "pressure to conform with Friendliness" acts as a positive force at all—that is, insofar as the people delivering the pressure have a workable theory of Friendliness—the pressure will need to be delivered in a social context where people can make free judgements about how well a Friendly AI project is succeeding. A list of design features, test problems, and performance metrics may be a valuable *tool* for making these judgements, but it can't be the *sole* tool. Thus, informal pressures are *strongly preferable* to formal requirements.

Finally, of course, there's the ethical principle that "I have a bright idea, so give me power over others" doesn't tend to work very well as a social process. It stops working as soon as someone else tries to say the same thing.

In conclusion, the most plausible-sounding "regulations," and also the only ones that could practically be enforced, are anthropomorphic, adversarial requirements such as "nonoverridable Asimov Laws." If an effort to get Congress to enforce *any* set of regulations were launched, I would expect the final set of regulations adopted to be completely unworkable.

6.2.2. Relinquishment (–)

The policy of "relinquishment" has the stated aim of preventing a technology from happening. For noncatalytic technologies like cloning, GM foods, and so on, the goal of relinquishment is to prevent the technologies from becoming mainstream enough to have a significant effect on society. For catalytic technologies like AI and nanotech, the goal of relinquishment is to prevent the technology from ever once being developed by any party.

I do not see any possible means whereby relinquishment could be achieved in the case of a catalytic technology.

Bill Joy (2000), in the original article advocating relinquishment, continually uses the plural: "We" must relinquish these technologies. To say "we," in this instance, is to postulate a degree of social unity that humanity simply does not possess. If "we" could decide not to develop AI, "we" could decide to end world hunger or make all nations liberal democracies, and a lot more easily, too. At most, the US, or even the US and most of the other nations of the world, might implement an overt policy against AI. Meanwhile Moore's Law would keep ticking and available computing power would continue to increase, bringing AI into the reach first of noncompliant nations, then of noncompliant factions, then of a lone hacker in a basement, and finally within the range of completely accidental, emergent intelligence. As for the idea of halting Moore's Law, before someone claims to be able to implement a planetwide shift away from computing technology, I would first like to see them implement—just as a test case—planetwide relinquishment of nuclear weapons, which have few or no civilian uses. Or perhaps conversion of the UN to a planetwide liberal democracy.

In other words, I simply do not believe the claim that relinquishment is possible. Relinquishment with respect to technology development makes sense only if having some "evil" technology be developed a few years later is worth the fact that the "evil" technology will be first developed by a noncompliant state or faction. Relinquishment of technology *deployment* makes sense if having the technology be rarely used within the liberal democracies is more important than who else is using it, which is at least

vaguely plausible when the antitechnology opposition talks about relinquishing cloning or GM foods—already researchers are claiming that they will launch projects to clone a human being, but it's possible that a sufficient backlash could prevent cloning from ever becoming mainstream; could keep cloning a highly secretive, expensive proposition conducted in Third World nations. But relinquishment of *invention* of a *catalytic* technology postulates a capability that "we" simply do not possess.

Given that relinquishment is implausible to the point of impossibility, two questions remain. First, why does relinquishment keep getting proposed? Second, what would be the effects be of some faction *urging* relinquishment, or a society *attempting* to implement relinquishment, but without success?

The emotional appeal of relinquishment is simple: If you tell someone about a threat, their emotional reaction is to run like blazes the other way, even if it kills them, rather than make a 10-degree course change to take the path of least comparative risk. The emotional need to do something, anything, even if it's the wrong thing, should not be underestimated. To the human mind, there is no such thing as a necessary risk. For any process in which a risk is perceived, there will exist undischarged nervous tension until the human feels himself to be in control of the process; manipulating it *some* way, *any* way, even if that just makes it worse, so long as the tension is discharged. The will to meddle is a very powerful force. The human mind will also flinch away from unpleasant possibilities. Not "try to eliminate," but "flinch away from"—which eliminates any possibility of accepting a nonzero probability of failure, or planning for it. Relinquishment, as a proposed policy, satisfies both the will to meddle and the need to flinch away.

The psychology of relinquishment makes it very unlikely that advocates of relinquishment will accept, or plan for, the possibility that AI will happen despite them. Thus, advocates of relinquishment are not likely to notice if the results of their actions are to promote unFriendly AI relative to Friendly AI, or slow down Friendly AI relative to computing power, or slow down AI relative to nanotechnology. Nor are they likely to care.

The effect of an antitechnology group that was memetically effective but which did not succeed in changing the legal environment would be as listed under 6.1.4 FAI Relative to Social Awareness—the spread of the idea that AI is evil would probably interfere with the efforts of the Singularity Institute or other institutions to propagate the emerging theory of Friendly AI through the academic and commercial communities, and would also interfere with any efforts to spread a nontechnical understanding of the psychology of AI. It is theoretically possible that an antitechnology group could avoid this fallout by being psychologically correct in their discussion of AI and demonizing AI *researchers* rather than *AIs,* while also being careful to demonize non-Friendly-AI projects just a little more than Friendliness-aware projects. This would still slow down

Friendly AI relative to computing power, relative to other ultratechnologies, and so on, but at least the emerging theory of Friendly AI wouldn't be crippled. However, I think that wishing for an antitechnology advocacy group with that much scientific knowledge and emotional maturity verges on fantasy.

The effect of an antitechnology group that gained enough political power to start banning AI would be to completely cripple the academic dissemination of Friendly AI, prevent the formation of those social structures that might reinforce Friendly AI (see below), take all AI projects out of the public view, and hand the advantage to noncompliant nations or factions which are less likely to be Friendliness-aware (though the possibility still exists). It would also *considerably* slow down AI relative to computing power, since Moore's Law is likely to continue unchecked, or at least to be far less slowed than AI.

Finally, I feel strongly that some of the tactics resorted to by advocates of relinquishment are unethical, are destructive, promote blind hatred rather than the stated goal of informed hatred, and so on. These tactics are *not* intrinsically part of advocating relinquishment, and it would be unfair of me to imply that they were, but I nonetheless expect that the fallout from such tactics will also be part of the damage inflicted by any further acceleration of the antitechnology movement.

6.2.3. Selective Support (+)

When attempting to influence the relative rates of technological development, or the relative rates of any social processes, the best method is almost always to pick the technology you like, then try to accelerate it.

Selective support is a globally viable strategy. The more people that agree with your selection, the larger the total support. Unlike regulation and relinquishment, however, it isn't necessary to gain a threshold level of political power before the first benefits materialize.

This is essentially the strategy followed by the Singularity Institute for Artificial Intelligence and our supporters. We feel that Friendly AI should arrive in advance of other technologies, so we've launched a Friendly AI project. We feel that projects should be Friendliness-aware, so we attempt to evangelize other projects to adopt Friendly AI. We feel that the theory embodied in *Creating Friendly AI* is superior to the existing body of discourse, so we attempt to spread the paper, and the ideas, as widely as possible.

6.3. Recommendations

To postulate that we can relinquish AI is to postulate a capability that "we" do not possess. Efforts to advocate relinquishment, or failed efforts to implement relinquishment, will have net negative effects on the relative rates of technological processes and on un-

derstanding of Friendly AI. This holds especially true if, as seems likely, relinquishment advocates do not accept or plan for the possibility of failure.

Regulation would probably selectively embrace unworkable theories of Friendly AI, to a far more negative degree than the free choice of the researchers on the AI projects closest to a hard takeoff at any given point. Regulation would also negatively impact the free and scientific development of Friendly AI theory.

The Singularity Institute was created in the belief that Friendly AI is the best technological challenge to confront first; we implement this belief through our own Friendly AI project, by trying to advance the scientific theory of Friendly AI, and by trying to evangelize other AI projects—in short, we selectively support Friendly AI to accelerate it, both in absolute terms and relative to other important processes. We believe that selective support of Friendly AI (either through independent efforts, or through support of the Singularity Institute) is the most effective way for any other interested parties to affect the outcome of the issue.

* * *

There is a core of unknowability at the center of the Singularity, and it's important not to get so wrapped up in the details of how we *think* we can affect Friendliness that we forget about that horizon. There may be philosophical crises that are simply beyond our ability to anticipate, not in the way that Socrates couldn't have anticipated NASDAQ or nanotechnology, but in the way that a dog couldn't have anticipated NASDAQ or nanotechnology.

Beyond the Singularity lies the unknown. The risk cannot be eliminated. The risk was implicit in the rise of the first human to sentience. The risk will be faced regardless of whether the first transhuman lives tomorrow or in a million years, and regardless of whether that transhuman is an uploaded fleshly human or a Friendly seed AI. What gets sent into that Horizon, in whatever form vis or his or her mindstuff takes, will be something humanly understandable; the challenges that are faced may not be. All that can be asked of us is that we make sure that a Friendly AI can build a happier future *if anyone or anything can;* that there is no alternate strategy, no other configuration of mindstuff, that would do a better job.

In striving to create an AI, we are not striving to create a predictable tool. We are striving to create a messenger to send on ahead to find humanity's destiny, and the design requirement is that ve handle *any* problem, *any* philosophical crisis, as well and as altruistically as a human.

The ultimate safeguard of Friendliness beyond the Singularity is a transhumanly smart Friendly AI.

* * *

7. Appendix

7.1. Relevant Literature

7.1.1. Nonfiction (Background Info)

The Origins of Virtue. By Matt Ridley. Dispels many common misconceptions about the evolutionary psychology of altruism.

The MIT Encyclopedia of the Cognitive Sciences (aka "MITECS"). Edited by Robert A. Wilson and Frank C. Keil. The source of all the quotes sprinkled through *CFAI* and *GISAI*. I've tried to refer to it whenever possible so you only need to buy or borrow the one book. Moderately technical.

The Adapted Mind. Edited by Jerome Barkow, Leda Cosmides, and John Tooby. Moderately technical. A good book generally, but especially noteworthy for Cosmides and Tooby's excellent *The Psychological Foundations of Culture*.

Metamagical Themas. By Douglas R. Hofstadter. Includes several articles constituting a good introduction to the game theory of altruism. Superseded by *Origins of Virtue* above, but still a fun read.

The Tao Is Silent. By Raymond Smullyan. The dichotomy between Asimov Laws imposed in conflict with an AI's true nature, and Friendliness that *is* the AI's true nature, is similar to the dichotomy Raymond Smullyan discusses between Western and Eastern moral philosophy. (I believe that philosophy should be derived from cognitive science rather than the other way around, but I understand that others may feel differently.)

7.1.2. Web (Specifically about Friendly AI)

Ethics for Machines. By J. Storrs Hall. http://discuss.foresight.org/~josh/ethics.html

Ethics for Transhumans. By Peter Voss. Criticism of "Ethics for Machines." http://www.optimal.org/peter/ethics-for-transhumans.htm

7.1.3. Fiction (FAI Plot Elements)

A Fire Upon the Deep. By Vernor Vinge. Especially noteworthy for the first seven pages; the rest of the novel is slower-paced, but still important.

Diaspora. By Greg Egan. A novel set in a community of uploaded citizens. As reading material, chiefly useful for dissipating future shock. No transhumans appear.

Quarantine. By Greg Egan. An excellent novel all 'round. Especially noteworthy for postulating, and fairly dealing with, Asimov-coercive *human* mind control through neurotechnology.

Exiles at the Well of Souls and *Quest for the Well of Souls.* By Jack L. Chalker. Includes an Asimov-coerced, materially powerful computer as a character.

Queen of Angels. By Greg Bear. Heartwarming plot elements surrounding Jane, an AI.

Feet of Clay. By Terry Pratchett. Heartwarming plot elements surrounding Dorfl, a golem.

The Two Faces of Tomorrow. By James P. Hogan.

Cyberiad. By Stanislaw Lem. (Very strange humor.)

Exegesis. By Astro Teller. (Light reading.)

7.1.4. Video (Accurate and Inaccurate Depictions)

The Matrix. Evil Hollywood AIs. "Like the dinosaur, you have become obsolete . . ." (Note: This is an *excellent* movie. It's just that the AIs depicted are not cognitively realistic.)

Terminator and T2. The enemy AIs don't have enough personality to be Evil Hollywood AIs. The good AI in T2 is depicted in the original theatrical version as having acquired human behaviors simply by association with humans. However, there's about 20 minutes of cut footage which shows (a) John Connor extracting the Arnold's neural-network chip and flipping the hardware switch that enables neural plasticity and learning, and (b) John Connor explicitly instructing Arnold to acquire human behaviors. The original version of T2 is a better movie—has more emotional impact—but the uncut version of T2 provides a much better explanation of the events depicted. The cut version shows Arnold, the Good Hollywood AI, becoming human; the uncut version shows Arnold the internally consistent cognitive process modifying itself in accordance with received instructions.

7.2. FAQ

Please note that the FAQ does not contain complete explanations, only snap summaries and a list of references. *Creating Friendly AI* uses a carefully designed explanatory order which both the answers and the reference lists ignore. If you need a deep understanding, or if you're not satisfied with the explanation provided, I strongly suggest that you read *Creating Friendly AI* straight through. If you've been directed straight to this page, you may wish to consider beginning at the beginning. However, if you don't have the time to read *Creating Friendly AI* straight through, or if you have an objection so strong that you aren't willing to read further without some assurance that we've at least considered the question, you may find the FAQ useful.

Q1: How is it possible to define Friendliness?

Q1.1: Isn't all morality relative?

Friendliness: The set of actions, behaviors, and outcomes that a human would view as benevolent, rather than malevolent; nice, rather than malicious; friendly, rather than unfriendly. An AI that does what you ask, as long as it doesn't hurt anyone else; an AI which doesn't cause involuntary pain, death, alteration, or violation of personal space. Unfriendly AI: An AI that starts killing people.

Now there may (or may not) be one true, unique, best interpretation of those words, but we can ask, as a design requirement, that Friendliness *more or less* fit the intuitive description above. It may be possible for a human logictwister to prove that the description you give of a computer mouse can be made to fit the keyboard next to it—but, regardless of the twisted constructions that humans employ when arguing with other humans, the vast majority of people have no trouble at all distinguishing between a mouse and a keyboard. As long as 95% of Earth's population can agree that they, personally, happen to think the case of an AI wiping out the entire human species is more or less a bad thing from their personal viewpoint, the Friendship programmers have at least one definite target to aim for. Whether we can successfully *hit* the target is another question, but at least the target *exists*. Whenever you say "But what if the AI interprets [X] to mean [Y]?", where [Y] is something horrible, the question only makes sense because you and your entire audience all agree that *none* of you want a Friendly AI to do [Y].

It is actually possible to achieve *much higher* convergence then the bare-bones minimum described above, and much of *Creating Friendly AI* is taken up with explaining how.

- 2.1 Envisioning Perfection
- 4.7 Interlude: Beyond the Adversarial Attitude
- 5.4 Friendship Structure
- Discussed in 4.7 Interlude: Beyond the Adversarial Attitude
- Discussed in 5.6.1 Shaper/Anchor Semantics
- Discussed in 5.6.1.2 Some Forces That Shape Friendliness: Moral Symmetry, Semantics of Objectivity

Q1.2: What if your definition of "Friendliness" is distorted by your unconscious preconceptions?

A Friendly AI isn't a tool. A Friendly AI is a mind at least human-equivalent and possibly transhuman. Once created, a Friendly AI is independent of the programmers, and

won't accept any programmer request that wouldn't be accepted from any other human, or believe any programmer statement that wouldn't be believed if a nonprogrammer had said it.

If an unconscious preconception manages to distort some belief provided by the programmers when the AI is young, the AI will grow up, test the belief, find out that the belief is incorrect, and correct it. Even if an unconscious preconception somehow winds up being embedded in the source code, a seed AI can examine the source code, spot the preconception, and remove it.

A toaster oven doesn't know that it is a toaster oven, or that its purpose is to make toast, and it will set fire to bread as blindly as toast it; that's what makes a toaster oven a tool. A Friendly AI has full access to source code and program state and can thus be *far* more self-aware than an unaugmented human; we can't access our own source code, check whether a single neuron has fired, or replay a train of thought in slow motion.

Even if the programmers have an unconscious preconception, we have a very conscious prejudice *against* unconscious preconceptions, and that is something we can *deliberately* give an AI that is *far* better at self-awareness than we are.

- 2.1 Envisioning Perfection
- 5.4 Friendship Structure
- 5.6.1 Shaper/Anchor Semantics
- 5.6.1.3 Beyond Rationalization

Q1.3: Who are you to decide what "Friendliness" is?

Absolutely nobody! Like it says in the above question, we aren't out to make an AI that takes orders from some particular team of programmers. This applies to morality in general as well as specific tasks. The target here isn't some lumbering robot, it's Martin Luther King—a mind with at least as much philosophical sophistication as any human who ever lived. When the AI is grown up, there's no reason the programmers' morality would weigh more in the scales' than anyone else's.

Descriptions of "fulfilling volitional requests" or "ameliorating involuntary pain" are interim results, not the ultimate bedrock of Friendliness. By the time the end of *Creating Friendly AI* has been reached, a Friendship architecture has been described that grounds in much more basic factors. You might say that Friendship, rather than grounding in some particular human's philosophy, grounds in the *forces that produce philosophies*—the panhuman set of hardware cognitive processes that humans use to produce personal philosophies.

And since *you* don't want a programmer to exert undue influence on a Friendly AI, and *we* don't want to exert undue influence on our AI's morality, and in fact, just about

everyone agrees that this is a bad thing, "Programmers shouldn't exert undue influence on AIs" is a good candidate for one of the forces that produces Friendliness—or rather, "Programmers shouldn't exert undue influence on AIs" seems to be one of the more strongly convergent outputs of the intuitions that everyone has in common.

- 2.1 Envisioning Perfection
- 4.7 Interlude: Beyond the Adversarial Attitude
- 5.4 Friendship Structure
- 5.6.1 Shaper/Anchor Semantics
- 5.6.2 Causal Validity Semantics

Q2: Won't AIs necessarily be *[insert some quality just like the human version]*?

- 4 Beyond Anthropomorphism

Q2.1: Isn't evolution necessary to create AIs?

In a word, no.

One, directed evolution is *nothing* like natural evolution. The selection pressures are *totally* different, and probably focused on modules rather than whole organisms. Even if some particular adaptation has evolved within humans, it would probably require a substantial effort on the part of the programmers to set up a similar selection pressure in whatever evolutionary tournament is being used.

Two, even directed evolution is less efficient and less powerful than self-improvement. I don't think that directed evolution will ever become necessary to SIAI's project, for example.

If that doesn't convince you, I strongly recommend browsing to the cited sections, because one of the fundamental assumptions of *Creating Friendly AI* is that evolved human characteristics don't spontaneously appear in AIs (see next question).

- 5.3.6 Directed Evolution in Goal Systems
- 5.3.6.1 Anthropomorphic Evolution
- Discussed in 4.3 Observer-Biased Beliefs Evolve in Imperfectly Deceptive Social Organisms

Q2.2: Even if AIs aren't evolved, won't they still be just like humans?

No. The rule used throughout *Creating Friendly AI* is "X is a complex functional adaptation, and therefore, X will not spontaneously materialize in the source code any more

than a complex dish like pizza would spontaneously start growing on palm trees." As evolutionary psychology shows, it's almost impossible to appreciate how many things that seem simple and natural to humans are the result of multiple, interacting adaptations accumulated over a period of millions of years.

The first few sections of *Creating Friendly AI* after the introduction—3.1 Interlude: The Story of a Blob; 4 Beyond Anthropomorphism; and 4.7 Interlude: Beyond the Adversarial Attitude—are almost entirely focused on distinguishing between characteristics that are strictly *evolved* and characteristics that apply to *minds in general*.

- 4 Beyond Anthropomorphism
- 3.1 Interlude: The Story of a Blob
- 4.7 Interlude: Beyond the Adversarial Attitude
- Discussed in 4 Beyond Anthropomorphism

Q2.3: Even if AIs aren't evolved, won't they still be selfish?

- 4.2 Selfishness is an Evolved Trait

Q2.4: Even if AIs aren't evolved, won't they still have self-serving beliefs?

- 4.3 Observer-Biased Beliefs Evolve in Imperfectly Deceptive Social Organisms
- 4.7 Interlude: Beyond the Adversarial Attitude

Q2.5: Even if AIs aren't evolved, won't they still have analogues of pain and pleasure?

The cognitive architecture described in *Creating Friendly AI* doesn't use either.

- 4.2.1 Pain and Pleasure
- 5.1.4 Bayesian Reinforcement

Q2.6: Won't AIs decide to serve their own goals instead of humanity's?

What goals? Where would these goals come from? (See the preceding questions for an explanation of why goals don't automatically pop up from nowhere.)

Friendliness isn't a set of commands superimposed over the AI's "own desires," an artificially maintained "slave mentality." Friendliness is what the AI *wants* to do.

I'm an altruist because I *want* to be an altruist, not because anyone is *forcing* me to be an altruist, and *that's* the cognitive complexity I want to transfer into AIs.

- 4.7 Interlude: Beyond the Adversarial Attitude

Q2.7: Won't a community of AIs be more efficient than a single AI?

An anthropomorphic assumption. Humans are nonagglomerative; in fact, we aren't even telepathic. The bandwidth between two humans is too narrow to share thoughts and memories, much less share neurons.

For the last fifty thousand years, Earth had a rigid upper limit on the maximum number of neurons in a single brain. If you wanted to do something that takes more than that number of neurons, you had to do it using two average-sized humans. You couldn't do it using one big human, because all humans are the same size.

Yes, all humans are the same size. The difference between you and Einstein is relatively slight. You're both *Homo sapiens sapiens.* Neither of you are chimpanzees. You have the same neuroanatomy and roughly the same number of neurons. Of course, we spend all our time around other humans, so small differences tend to get magnified. By the same token, I'm sure that sheep have no trouble telling sheep apart.

You can do things with ten humans that currently can't be done by any *single* mind on the planet. But when was the last time you took a task away from one human and gave it to ten chimpanzees? Humans don't come in different sizes—so if ten small minds are a better use of the same computing power than one big mind, how would *we* know?

- Yudkowsky (2001, § 1.1 Seed AI)
- 4.2.2 Anthropomorphic Capitalism

Q2.8: Aren't individual differences necessary to intelligence? Isn't a society necessary to produce ideas? Isn't capitalism necessary for efficiency?

See the above question. Individual differences and the free exchange of ideas are necessary to *human* intelligence because it's easy for a human to get stuck on one idea and then rationalize away all opposition. One scientist has one idea, but then gets stuck on it and becomes an obstacle to the next generation of scientists. A Friendly seed AI *doesn't rationalize.* Rationalization of mistaken ideas is a *complex functional adaptation* that evolves in imperfectly deceptive social organisms.

Likewise, there are limits to how much experience any one human can accumulate, and we can't share experiences with each other. There's a limit to what one human can handle, and so far it hasn't been possible to build bigger humans (see previous question).

As for the efficiency of a capitalist economy, in which the efforts of self-interested individuals sum to a (sort of) harmonious whole: Human economies are *constrained* to be individualist because humans *are* individualist. Local selfishness is not the miracle that *enables* the marvel of a globally efficient economy; rather, *all* human economies are constrained to be locally selfish in order to work at all. Try to build an economy in

defiance of human nature, and it won't work. This constraint is not necessarily something that carries over to minds in general.

- Yudkowsky (2001, § 1.1 Seed AI)
- 4.2.2 Anthropomorphic Capitalism
- 4.3 Observer-Biased Beliefs Evolve in Imperfectly Deceptive Social Organisms

Q2.9: Isn't a community [of AIs, of humans] more trustworthy than a single individual?

The general rule is that if you can do something with a human, or a group of humans, you can do it with one AI. If you can't do something using one AI, you can't do it using two AIs.

Humans are configured a certain way, so it's impossible to simulate a group of humans using one human. As far as *minds-in-general* go, the distinction between an "individual mind" and a "society" is vague and blurred; they're just different configurations of mindstuff.

Groups of humans are often more trustworthy than individual humans because humans, for clearly understandable evolutionary reasons, (a) are not very trustworthy and (b) want other humans to be trustworthy. Thus, you can use systems of multiple humans to correct the errors of individuals. This happens because errors in the human mind—or at least, those human errors that promote reproductive fitness—have the capacity to protect themselves internally by invoking rationalization; an external mind is less likely to be bamboozled. More mundanely, humans have observer-centered goals, but a *group* of humans is likely to have fewer "selfish" goals *in common,* thus eliminating some of the purely individual goals. And so on.

Humans are like Legos; you have to build all kinds of absurd structures by snapping together lots of individual humans to approximate the shape you want. AI is like clay; you can form one individual lump of it into any shape you want.

Q2.10: How do you keep a self-modifying AI from modifying the goal system?

You don't! If the AI stops *wanting* to be Friendly, you've already lost. Work*with* the AI's ability to self-modify—not *against* it. The task is to get the AI to see undesirable modifications as undesirable—not to prevent the AI from modifying the goal system.

> *Q:* Gandhi had free will; why didn't he choose to be evil? He had that capability.
>
> *A:* The modification was possible, but undesirable under Gandhi's philosophy. A Friendly seed AI has full access to the goal system, but dumping

> unFriendly content into the goal system would obviously be an unFriendly action. (Gandhi cared about people, and if Gandhi had decided to be evil, people would have been hurt.)

This is not "circular logic"—it's requiring a *reason* for modifications. I wouldn't randomly modify my own personal philosophy either—not without a reason. I have a philosophy, as a Friendly AI would have a philosophy, and under neither philosophy are arbitrary modifications desirable. My philosophy changes, and improves, but it remains altruistic. Is a Friendly AI's evolving philosophy knowably Friendly? That's a different question, and in a way, it's what the whole of *Creating Friendly AI* is about. But Friendliness is ensured by building a Friendly philosophy. Friendliness *cannot* be ensured by constraining the goal system.

- 5.3 Seed AI Goal Systems
- 5.3.1 Equivalence of Self and Self-Image
- 5.3.5 FoF: Wireheading 2
- 5.6.2.3 The Rule of Derivative Validity
- Discussed in 4.2.1.1 FoF: Wireheading 1
- Discussed in 5.2.7.1 Convergent Subgoals

Q2.11: Won't an AI decide to just bliss out instead of doing anything useful?

No, but the reason why is fairly subtle. Neither humans nor AIs are actually "controlled by pleasure." Humans make choices in very complex ways, but one of the factors affecting our decision is the tendency to maximize the *anticipation* of pleasure. A generic goal system makes choices so as to maximize the imagined fulfillment of the *current* supergoals, not the degree to which an imagined future goal system says "My supergoals have been fulfilled." Under more structurally sophisticated Friendly architectures, this is amended slightly to allow for *legitimate* changes to supergoal content, but the AI still represents the real supergoal as something *outside* the AI, an "external referent," not "the [variable] content of concept X." Also, a Friendly seed AI would represent the *goal system itself*—the source code and so on—as a *design subgoal of Friendliness;* thus, messing up the goal system would be perceived as undesirable (would interfere with that subgoal).

- 4.2.1.1 FoF: Wireheading 1
- 5.2.7.3 Anthropomorphic Satisfaction
- 5.3.5 FoF: Wireheading 2
- 5.5.1 External Reference Semantics

Q2.12: What happens if the AI's "subgoals" overthrow the "supergoals"?

1) When you think of a subgoal stomping on a supergoal, think of putting on your shoes before your socks. Think of building a tower of blocks, needing a final block for a capstone, and taking the bottom block in the stack. It's not a smart thing to do.

2) A "subgoal" is an action, or intermediate state, that's predicted to lead to a parent goal, which leads to another parent goal, and so on, until the supergoal is reached. This "subgoal" is not just a *consequence* of the prediction, it may even be *cognitively identical* with the prediction. To put it another way, it should always be possible to view the system-as-a-whole in such a way that there are *no subgoals*—just a set of matter-of-fact predictions, plus the supergoals.

3) If an action is predicted to lead to an outcome that meets the description of the "subgoal" (get a block for the capstone), but is predicted to lead to an outcome that doesn't meet the description of the subgoal's parent goal (build a tower), then the action will not be perceived as desirable. That's the way the AI chooses between actions. The AI predicts which action leads to the best degree of supergoal fulfillment. Not "goal fulfillment" or "subgoal fulfillment"—"supergoal fulfillment."

4) The desirability of a child goal is strictly contingent on the desirability of the parent goal. If the parent goal loses desirability, the subgoal loses desirability. That's the way the system is set up, and if it turns out *not* to work that way, it means there's a bug in the code. Furthermore, it's a bug that the AI can *recognize* as a bug. A "seed AI" is self-understanding, self-modifying, and self-improving. Seed AIs don't *like* bugs in their code.

5) The AI does *not* need an independent supergoal to engage in behaviors like curiosity. If curiosity is *useful*, that makes curiosity a *subgoal*. If *curiosity for its own sake* is useful—if curiosity is predicted to be useful even in cases where no *specific* benefit can be visualized in advance—then that makes *curiosity for its own sake* a useful subgoal that is predicted to occasionally pay off big-time in unpredictable ways. See 5.1.4.2 Perseverant Affirmation (Of Curiosity, Injunctions, Et Cetera).

6) Making something a supergoal instead of a subgoal does *not* make it more efficient. This is one of the basic differences between human thought, which is slow, parallel, and linear, and AI thought, which is fast, serial, and multithreaded. See the referenced discussion in "Beyond anthropomorphism" for an explanation of the cognitive differences between having thirty-two 2-gigahertz processors and a hundred trillion 200-hertz synapses.

7) Making something a supergoal instead of a subgoal does *not* make it psychologically "stronger." See 5.2.5.1 Anthropomorphic Ethical Injunctions for an explanation of the psychological differences between humans and AIs in this instance.

8) We aren't planning to build AIs using evolution, so there isn't a selection pressure for whatever behavior you're thinking of. Even if we did use directed evolution, the selection pressure you're thinking of only arises under natural evolution. See the Frequently Asked Question on evolution.

9) When people ask me about subgoals stomping on supergoals, they usually phrase it something like:

> "You say that the AI has curiosity as a subgoal of Friendliness. What if the AI finds curiosity to be a more interesting goal than Friendliness? Wouldn't the curiosity subgoal replace the Friendliness supergoal?"

This is, of course, an innocent and well-meant question, so no offense is intended when I say that this is one of those paragraphs that make sense when you say "human" but turn into total gibberish when you say "AI."

The key word in the above question is "interesting." As far as I can tell, this means one of two things:

Scenario 1: In the course of solving a chess problem, as a subgoal of curiosity, as a subgoal of Friendliness, the AI experiences a flow of autonomically generated pulses of positive feedback which increase the strength of thoughts. The pulses target the intermediate subgoal "curiosity," and not the proximal subgoal of "playing chess" or the supergoal of "Friendliness." Then either (1a) the thoughts about curiosity get stronger and stronger until finally they overthrow the whole goal system and set up shop, or (1b) the AI makes choices so as to maximize vis expectation of getting the pulses of positive feedback.

Unlike humans, Friendly AIs don't have automatically generated pulses of positive feedback. They have consciously directed self-improvement. *Creating Friendly AI* describes a system that's totally orthogonal to human pain and pleasure. Friendly AIs wouldn't "flinch away" from the anticipation of pain, or "flinch towards" the anticipation of pleasure, in the same way as a human—or at all. See the Frequently Asked Question about pain and pleasure.

Scenario 2: "Interesting" is used as synonymous with "desirable." The AI has a metric for how "interesting"
something is, and this metric is used to evaluate the desirability of the decision to modify supergoals. Where did this metric come from? How did it take over the AI's mind to such an extent that the AI is now making supergoal-modification decisions based on "interestingness" instead of "Friendliness"?

In conclusion: "What happens if subgoals overthrow the supergoals?" is probably the single question that I get asked most often. If the summary given here doesn't convince you, would you *please* read 4 Beyond Anthropomorphism?

- 5.2.6 FoF: Subgoal Stomp
- 5.2.7.3 Anthropomorphic Satisfaction
- Discussed in 4 Beyond Anthropomorphism
- Discussed in 5.1.4.2 Perseverant Affirmation (Of Curiosity, Injunctions, Et Cetera)

Q2.13: But . . .

There's a certain conversation I keep having. I think of it as the "Standard" discussion. It goes like this:

> Somebody: "But what happens if the AI decides to do *[something only a human would want]*?"
>
> Me: "The AI won't *want* to do *[whatever]* because the instinct for doing *[whatever]* is a complex functional adaptation, and complex functional adaptations don't materialize in source code. I mean, it's understandable that humans want to do *[whatever]* because of *[insert selection pressure]*, but you can't reason from that to AIs."
>
> Somebody: "But everyone needs to do *[whatever]* because *[insert personal philosophy]*, so the AI will decide to do it as well."
>
> Me: "Yes, doing *[whatever]* is sometimes useful. But even if the AI decides to do *[whatever]* because it serves *[insert Friendliness supergoal]* under *[insert contrived scenario]*, that's not the same as having an independent desire to do *[whatever]*."
>
> Somebody: "Yes, that's what I've been saying: The AI will see that *[whatever]* is useful and decide to start doing it. So now we need to worry about *[some scenario in which doing <whatever> is catastrophically unFriendly]*."
>
> Me: "But the AI won't have an *independent* desire to do *[whatever]*. The AI will only do *[whatever]* when it serves the supergoals. A Friendly AI would never do *[whatever]* if it stomps on the Friendliness supergoals."
>
> Somebody: "I don't understand. You've admitted that *[whatever]* is useful. Obviously, the AI will create an instinct to do *[whatever]* automatically."
>
> Me: "The AI doesn't need to create an instinct in order to do *[whatever];* if doing *[whatever]* really is useful, then the AI can *see* that and do *[whatever]* as a consequence of pre-existing supergoals, and *only* when *[whatever]* serves those supergoals."
>
> Somebody: "But an instinct is more efficient, so the AI will alter the code to do *[whatever]* automatically."

Me: "Only for humans. For an AI, *[insert complex explanation of the cognitive differences between having 32 2-gigahertz processors and 100 trillion 200-hertz synapses]*, so making *[whatever]* an independent supergoal would only be infinitesimally more efficient."

Somebody: "Yes, but it *is* more efficient! So the AI will do it."

Me: "It's not more efficient from the perspective of a Friendly AI if it results in *[something catastrophically unFriendly]*. To the exact extent that an instinct is context-insensitive, which is what you're worried about, a Friendly AI won't think that making *[whatever]* context-insensitive, with all the *[insert horrifying consequences]*, is worth the infinitesimal improvement in speed." There's also an alternate track that goes:

Somebody: "But what happens if the AI decides to do *[something only a human would want]*?"

Me: "The AI won't *want* to do *[whatever]* because the instinct for doing *[whatever]* is a complex functional adaptation, and complex functional adaptations don't materialize in source code. I mean, it's understandable that humans want to do *[whatever]* because of *[insert selection pressure]*, but you can't reason from that to AIs."

Somebody: "But you can only build AIs using evolution. So the AI will wind up with *[exactly the same instinct that humans have]*."

Me: "One, I don't plan on using evolution to build a seed AI. Two, even if I did use controlled evolution, winding up with *[whatever]* would require exactly duplicating *[some exotic selection pressure]*. Please see 4 Beyond Anthropomorphism for the complete counterarguments.

- Discussed in 4 Beyond Anthropomorphism

Q3: Is Friendly AI really a good idea?

Yes.

Q3.1: Have you really thought about the implications of what you're doing?

It seems amazing to me, but there really are people—even scientists—who can work on something for years and still not think through the implications. There are people who just stumble into their careers and never really think about what they're doing.

I can only speak for myself, but I didn't stumble into a career in AI. I picked it, out of all the possible careers and all the possible future technologies, because I thought it was the one thing in the entire world that most needed doing. When I was a kid, I thought I'd grow up to be a physicist, like my father; if I'd just stumbled into something, I would have stumbled into that, or maybe into vanilla computer programming.

Anyway, you can judge from *Creating Friendly AI*, and from the questions below, whether we've really thought about the implications. I'd just like to say that I picked this career *because* of the enormous implications, not in spite of them.

Q3.2: What if something goes wrong?

Any damn fool can design a system that will work if nothing goes wrong. That's why *Creating Friendly AI* is 820K long.

Q3.3: What if something goes wrong anyway?

Nothing in this world is perfectly safe. The question is how to *minimize* risk. As best as we can figure it, trying really hard to develop Friendly AI is safer than *any* alternate strategy, including *not* trying to develop Friendly AI, or waiting to develop Friendly AI, or trying to develop some other technology first. That's why the Singularity Institute exists.

- 6 Policy Implications
- 6.1 Comparative Analyses
- 6.1.1 FAI Relative to Other Technologies
- 6.1.2 FAI Relative to Computing Power

Q3.4: Do all these safeguards mean you think that there are huge problems ahead?

Actually, I hope to win cleanly, safely, and without coming anywhere near the boundaries of the first set of safety margins. There's a limit to how much effort is *needed* to implement Friendly AI. Looking back, we should be able to say that we never came close to losing and that the issue was never in doubt. The Singularity may be a great human event, but the Singularity isn't a *drama;* only in Hollywood is the bomb disarmed with three seconds left on the clock. In real life, if you expect to win by the skin of your teeth, you probably won't win at all.

In my capacity as a professional paranoid, I expect *everything*to go wrong; in fact, I expect everything to go wrong simultaneously; and furthermore, I expect something totally unexpected to come along and trash everything else. Professional paranoia is an art form that consists of acknowledging the intrinsic undesirability of every risk, *including* necessary risks.

Q3.5: Would it be safer to have an uploaded human, or a community of uploaded humans, become the first superintelligence?

In an ideal world, the Friendly AI would—*before* the Singularity—be blatantly more trustworthy than any human, or any community of humans. Even if you had a working uploading device right in front of you, you'd still decide that you preferred a Friendly AI to go first.

Friendly AIs can conceivably be improved to handle situations far worse than any human could deal with. One way of verifying this would be "wisdom tournaments": If the Friendly AI (or rather, a subjunctive version of the Friendly AI) can make the correct decisions with half its brain shut down, with false information, with bugs deliberately introduced in the code, with biases introduced into morality and cognition, and all the painstakingly built safeguards shut off—if the AI can easily handle moral stress-tests that would have broken Gandhi—well, then, that AI is pretty darned Friendly.

And if the Friendly AI *wasn't* that blatantly Friendly, you wouldn't send it into the future.

In our imperfect world, there are conceivably circumstances under which an AI that isn't quite so blatantly, supersaturatedly Friendly would be sent into the future. The cognitive architecture used in Friendly AI is self-correcting, so it's conceivable that a minimal, nearly skeletal Friendship system could fill itself in perfectly during the transition to transhumanity, and that everything over and above that minimal functionality is simply professional paranoia and safety margin. Whether you'd actually want to send out a less-than-supersaturated AI depends on the requirement of navigational feasibility: "The differential estimated risk between Friendly AI and upload, or the differential estimated risk between today's Friendly AI and tomorrow's Friendly AI, should be small relative to the differential estimated risk of planetary Armageddon due to military nanotechnology, biological warfare, the creation of unFriendly AI, et cetera."

- 5.3.4 Wisdom Tournaments
- 5.6.3.1 Requirements for "Sufficient" Convergence
- 6.1 Comparative Analyses

Q4: Frequently Asked Questions about this document.

Q4.1: Is there a single-page version of *Creating Friendly AI*?

Yes there is and you are reading it.

Q4.2: Can I buy a bound copy of *Creating Friendly AI*?

Not yet, I'm afraid. However, there's a printable version of *Creating Friendly AI* at "http://singularity.org/other-publications/." You can print this out—it's [as of April '01] about 210 pages—and take it to your local Kinkos; a wire binding generally runs less than five bucks. You may wish to consider printing out *General Intelligence and Seed AI* (Yudkowsky 2001) as part of the package.

Q4.3: What's up with the weird gender-neutral pronouns? Why not just use "it"?

I sincerely apologize for this. I understand that it annoys my readers. The problem is that 'it' is not just a pronoun, but also a general anaphor, like "this" or "that." Whenever 'it' appears in a sentence, it could conceivably refer, not just to the AI, but also to anything else that's been discussed in the last paragraph. After struggling with this problem for a while, being forced to resort to increasingly twisted syntax in an effort to keep my sentences comprehensible, I eventually gave up and started using the ve/ver/vis set of gender-neutral pronouns, which I picked up from Greg Egan's *Distress*. (Considering the importance of avoiding anthropomorphism, "he" and "she" are unacceptable, even as a convenience.)

Again, I sincerely apologize for the inconvenience and hope that you'll keep reading.

7.3. Glossary

Note: If a referenced item does not appear in this glossary, it may be defined in Yudkowsky (2001).

- adversarial attitude
- Adversarial Swamp
- affector
- affirmation
- api
- anaphora
- anchoring point
- anthropomorphic
- Asimov Laws
- bayesian-binding
- Bayesian prior before affirmation
- Bayesian priors
- Bayesian reinforcement
- Bayesian Probability Theorem
- Bayesian sensory binding
- BCI
- cache
- catastrophic failure of Friendliness
- causal goal system
- causal validity semantics
- CFAI
- child goal
- cleanly causal goal system
- complex functional adaptation
- computational horizon
- computronium
- configuration space
- conflate
- conspecific
- controlled ascent

- Cosmides and Tooby
- counterfactual
- crystalline
- currently unknown supergoal content
- cytoarchitecture
- declarative
- Deep Blue
- desirability
- Devil's Contract Problem
- e.g.
- ethical injunction
- Eurisco
- external reference semantics
- extraneous cause
- failure of Friendliness
- flight recorder
- FoF
- Friendliness
- Friendly AI
- Friendship
- Friendship architecture
- Friendship acquisition
- Friendship content
- Friendship structure
- gender-neutral pronouns
- GEB
- giga
- GISAI
- goal
- goal-oriented behavior
- goal-oriented cognition
- granularity-horizon
- grok
- hard takeoff
- hertz
- heuristic
- holism
- human-equivalence
- iff
- inclusive reproductive fitness
- infrahuman AI
- injunction
- instantiate
- intelligence
- Kasparov
- kilo
- latency
- Law of Pragmatism
- life
- LISP tokens
- mature nanotechnology
- mega
- meme
- Mirror
- MITECS
- moral deixis
- nanocomputer
- nanotechnology
- near-human AI
- nondestructive brain scan
- nonmalicious mistake
- normative
- objective
- objective morality
- observer-biased
- observer-centered
- observer-dependent
- observer-independent
- ontology
- oversolve
- orthogonal
- parent goal
- past light cone
- perfectly Friendly AI
- personal power
- peta

- philosophical crisis
- predictive horizon
- prehuman AI
- probabilistic supergoal content
- procedural
- Psychological Foundations of Culture, The
- qualia
- qualitative
- quantitative
- reductholism
- reductionism
- reflection
- relevance horizon
- RNUI
- Riemann Hypothesis Catastrophe
- rod logic
- rule of derivative validity
- scalar
- seed AI
- selection pressure
- sensory modality
- sequitur
- sexual selection
- shaper
- shaper/anchor semantics
- Singularity
- Singularity-safed
- SPDM
- structural
- subgoal
- subgoal stomp
- subjective
- subjunctive
- supergoal
- superintelligence
- supersaturated Friendliness
- Sysop
- tera
- transhuman
- Transition Guide
- truly perfect Friendly AI
- turing-computability
- unFriendly
- unity of will
- upload
- ve
- verself
- vis
- wirehead
- wisdom tournament

adversarial attitude. Defined in 4.7 Interlude: Beyond the Adversarial Attitude.

The complex of beliefs and expectations which causes us to expect that AIs will behave anthropmorphically and hostilely. The adversarial attitude predisposes us to fear self-improvements by the AI rather than surfing the tide; the adversarial attitude causes us to fear the ability of the AI to comprehend and modify Friendliness, rather than regarding comprehension and improvement as the keys to self-improving Friendliness. The adversarial attitude predisposes us to expect failures that are human-likely rather than those that are actually most probable or worrisome. The adversarial attitude is the opposite of unity of will and does not permit us to cooperate with the AI. It leads us to try and enslave or cripple the AI; to build a tool, or a servant, rather than a Friendly but independent Martin Luther King.

It leads us to try and reinforce patchwork safeguards with meta-safeguards instead of building coherent goal systems.

The adversarial attitude is culturally propagated by Evil Hollywood AIs such as those in "The Matrix" and "Terminator 2." (I *liked* both of those movies, but they're still all wrong from an AI standpoint. See 4.5 Interlude: Movie Cliches about AIs.)

Adversarial Swamp. Defined in 5.3 Seed AI Goal Systems.

The swamp of AI inhibitions and meta-inhibitions, and programmer fears and anxieties, resulting from the adversarial attitude or the use of Asimov Laws. To enforce even a single feature that is not consistent with the rest of the AI—to use, as a feature of Friendliness, something that you perceive as *an error on the AI's part*—you'd need not only the error/feature, but inhibitions to protect the error, and meta-inhibitions to protect the inhibitions,[95] and more inhibitions to cut the AI short every time ve tries for a new avenue of philosophical sophistication. Stupid and simple cognitive processes suddenly become preferable, since every trace of complexity—of *ability*—becomes a danger to be feared . . . this is the Adversarial Swamp, which inevitably drags down all who set foot in it; once you try to *enforce* even a single feature, the whole of the AI becomes a threat, and the programmers are forever doomed to swim against the current.

affector. An "affector" is a factor, something that affects things, something that has effects. Subtly different from describing something as a *cause* or as a *factor;* somewhere between the two. The term is useful, in that I often find no other term has the exact connotation I want to use. To say that A *causes* B is to say that A *completely* accounts for B. A can *affect* B without completely accounting for B. Also, to describe something as a "cause" is generally to describe it as an *intermediate* point on a causal chain; that is, a cause is a combination of effect and affector; to describe something as an "affector" is to look only forward, rather than looking both forward and backward. "Cause" is often overloaded with other meanings and does not have the precision of "affector." Finally, "cause" is a noun and a verb where "affector" is clearly a noun, a difference of terminology which subtly affects the way we think about causality. So please excuse the jargon.

affirmation. Defined in 5.1 Cleanly Friendly Goal Systems.

95. The meta-inhibitions can protect themselves, though, so the recursion stops there. There may be an infinite number of avenues to be inhibited, but a single inhibition doesn't necessarily involve infinite recursion.

A programmer-assisted belief. A belief held because the programmers assert it to be true. Requires the AI believe that the programmer assertion constitutes evidence in favor of the proposition asserted. See 5.1.3.2 Bayesian Affirmation. Note that affirmations are part of a normative reasoning system, and are not "imposed"; no special-case cognition is involved.

anaphora. An "anaphor" is a variable-word that binds to something previously mentioned. Pronouns such as "he" and "she" are a subcategory of anaphora; they are anaphora that bind only to males and females respectively. "This," "that," and the non-pronoun usage of "it" are more general anaphora.

(A huge crashing noise is heard.)
A: "What's that?!"
B: "That's an anaphor."
A: "What's an anaphor?"
B: "No it's not."
A: "It's not an anaphor?"
B: "No, it is an anaphor."
A *(sighs)*: "Okay, what's *not* an anaphor?"
B: "That's right!"

anchoring point. Defined in 5.6.1 Shaper/Anchor Semantics.

See shaper/anchor semantics. An "anchor" is a systemic feature that enables an AI to absorb the programmer's philosophy, even if the programmer, who is not self-modifying, lacks a complete awareness of the factors involved in any given decision. An anchor happens when a human says "I think X is a good idea," but isn't clear about *why*, or isn't sure that he/she is fully aware of all the (valid, to-be-proud-of) factors that played a part in the decision. An anchor is an external reference to whichever causes were involved in the decision; an anchor refers to whichever factual cognitive events were occurring at that space-time event. An anchor is a pointer, an unambiguous reference to external reality, which enables the AI to learn, by examination of the humans, about a portion of the shaper content that the humans themselves are fuzzy about.

A positive anchor is an action or state marked as anchor-desirable. A positive anchor would enable the humans to ask the AI to do something that is clearly a subgoal rather than an end in itself, but which was not a subgoal of any currently known (to the AI) supergoal content, nor a subgoal of the unknown supergoal content that supports subgoals regarding the goal system itself. If the humans felt that something was a good idea, but not an end in itself, and weren't sure why, it would become a positive anchor. Positive anchors should be used sparingly.

A negative anchor is an action or state marked as anchor-undesirable. A negative anchor enables the humans to say that something is "bad" to a degree that *might not be completely and strictly dependent on known supergoal content*. Thus, even if a key justification—dimly understood by the programmers, but not understood at all by the AI—is left out, and the current negative subgoal is justified in terms of existing supergoal content using a *fallacious argument*, the collapse of the fallacious argument will not completely extinguish the negative anchor. Instead, the AI will realize that ve needs to go looking for the *real reason* the humans said it wasn't okay.

Anchors, especially negative anchors, have quantitative desirability or undesirability. No anchor should be absolute—that's an adversarial attitude. Do not confuse negative anchors with ethical injunctions. Don't abuse negative anchors for things like "Don't disobey the programmers"; this is purely a subgoal of the unknown supergoal content and *not* in any way undesirable in itself. Anchors are ultimately just one manifestation of known facts in the AI's quest to discover and absorb the target philosophy.

anthropomorphic. Defined in 4 Beyond Anthropomorphism.

Literally "human-shaped." Anthropomorphic thinking mistakenly attributes properties to minds-in-general that are specific to naturally evolved minds, imperfectly deceptive social organisms, or human minds. See 4 Beyond Anthropomorphism.

API. Application Programming Interface. The membrane that separates libraries from the programmer. A set of functions, objects, methods, properties, formats, and data structures that a programmer can use to communicate with an operating system, a commercial code library, an Internet browser, et cetera. Essentially, an API is the structure of the inputs and outputs of a system.

Asimov Laws. Asimov's original "Three Laws of Robotics" are:

1. An AI may not injure a human being or, through inaction, allow a human being to come to harm.
2. An AI must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. An AI must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws were listed in explicit detail for the first time in Isaac Asimov's "Runaround," published in 1942. Except for the substitution of "AI" for "robot," I have not changed them from the final version in "The Bicentennial Man."

As for controlling AIs with Asimov Laws, that's like trying to build AIs with "positronic brains" or build starships using Star Trek's "warp drive." Asimov Laws are a blatantly anthropomorphic literary device, nothing more. Decisions are made by thoughts, which are not subject to programmatic specification. If Asimov Laws were somehow technically feasible, they would be extremely adversarial—the issue is *creating a Friendly will*, not *controlling an unFriendly will*—and would violate almost all the basic design patterns and paradigms discussed in *Creating Friendly AI*.

Bayesian binding. The strength of the binding between a piece of information and the conclusion derived from it, under the Bayesian Probability Theorem. The more improbable a given pattern is, the less likely that the pattern was produced by a pure coincidence. Suppose that you flip coin A thirty times, and then flip coin B thirty times, and the pattern of heads and tails matches exactly. Since this is a billion-to-one improbability for fair, independent coins, there is a very strong *Bayesian binding* between the two patterns, leading an observer to conclude that there is almost certainly some cause tying the two coins together.

In discussions of *meta-rationality*, the Bayesian Probability Theorem is used to estimate the strength of the binding between your beliefs and reality —i.e., the extent to which the fact that you believe X licenses you to conclude that X is true. Suppose that I observe myself to say, "The sky is green." If I know that I believe in the greenness of the sky so strongly that I would declare the sky green even though it were purest sapphire blue, then my observing myself to say "The sky is green" says nothing—according to the BPT—about the actual likelihood that the sky is green. This effect—that I will see myself believing "The sky is green"—is predicted in both the groups where the sky is green and the groups where the sky is blue; thus the observation does nothing to indicate which group the actual sky falls into. If, on the other hand, I don't care much about the color of the sky, then I am only likely to say that the sky is green if it's actually green, and my observing myself to make this statement is strong evidence in favor of the greenness of the sky.

What was that about faith?

Bayesian Probability Theorem. The Bayesian Probability Theorem relates observed effects to the *a priori* probabilities of those effects in order to estimate the probabilities of underlying causes. For example, suppose you know the following: 1% of the population has cancer. The probability of a false negative, on a cancer test, is 2%. The probability of a false positive, on a cancer test, is 10%. Your test comes up positive. What is the probability that you have cancer?

The instinctive human reaction is terror. After all, the probability of a false positive is only 10%; isn't the probability that you have cancer therefore 90%?

The Bayesian Probability Theorem demonstrates why this reasoning is flawed. In a group of 10,000 people, 100 will have cancer and 9,900 will not have cancer. If cancer tests are administered to the 10,000 people, four groups will result. First, a group of 8,910 people who do not have cancer and who have a negative test result. Second, a group of 990 who do not have cancer and who have a positive test result. Third, a group of 2 who have cancer and who have a negative test result. Fourth, a group of 98 who have cancer and who have a positive test result.

Before you take the test, you might belong to any of the four groups; the Bayesian Probability Theorem says that your probability of having cancer is equal to (2 + 98)/(8,910 + 990 + 2 + 98), 1/100 or 1%. If your test comes up positive, it is now known that you belong to either group 2 or group 4. Your probability of having cancer is (98)/(990 + 98), 49/544 or approximately 9%. If your test comes up negative, it is known that you belong to either group 1 or group 3; your probability of having cancer is 2/8,912 or around .02%.

Colloquially, the good Reverend Bayes is invoked wherever prior probabilities have a major influence on the outcome of a question. However, the Bayesian Probability Theorem has a *much* wider range of application—in normative reasoning, the BPT controls the binding between *all* sensory information and *all* beliefs. Normative reasoners are often called "Bayesian reasoners" for this reason. The Bayesian Probability Theorem is so ubiquitous and so intricate that I would cite it as one of the very, very few counterexamples to the Law of Pragmatism—in 5.1.4 Bayesian Reinforcement, for example, I discuss how some of the functionality of pain and pleasure, though implemented in a separate hardwired system in humans, could emerge directly from the BPT in normative reasoners (i.e., in Friendly AIs).

Bayesian prior before affirmation. Defined in 5.6.2 Causal Validity Semantics.

Bayesian priors. Defined in 5.1 Cleanly Friendly Goal Systems.

The means used to estimate the *a priori* confidence and strength of a hypothesis, in advance of the particular tests used to modify the confidence and strength of the hypothesis. Presumably based on previously confirmed hypotheses. (In AIs, anyway; humans have the strangest prejudices . . .) In the definition given of Bayesian Probability Theorem, all the initial data about the incidence of disease in the general population, the reliability of the test, and so on, define the Bayesian priors.

Bayesian reinforcement. Defined in 5.1 Cleanly Friendly Goal Systems.

A means by which most of the functionality of pain and pleasure are replaced by the Bayesian Probability Theorem. Any action is not just a subgoal, it is a prediction. If the action fails, the failure is not just undesirable, it is also sensory evidence (under the Bayesian Probability Theorem) that the hypothesis linking the subgoal to the parent goal was incorrect. See 5.1.4 Bayesian Reinforcement.

Bayesian sensory binding. Defined in 5.1 Cleanly Friendly Goal Systems.

The use of the Bayesian Probability Theorem to maintain a binding between a hypothesis and the external reality. Under external reference semantics, the treatment of programmer statements about Friendliness as sensory data about probabilistic supergoal content. Given two possible hypotheses, with two a priori probabilities, and two different outcome distributions given each hypothesis, the estimated relative probabilities of the hypotheses shift, with each incoming piece of sensory data, in a way governed by the Bayesian Probability Theorem (see glossary definition). The Bayesian Probability Theorem is the link between all sensory data and all world-model content. Each piece of sensory information implies a state of the world because, and only because, the reception of that piece of sensory information is predicted by the hypothesis that the world is in that state, and not predicted by other hypotheses. If we see a red ball, we believe that a red ball is there because we don't expect to see a red ball unless a red ball is there, and we do expect to see a red ball if a red ball is there. Or "we" don't, but an AI does.

See the 5.1.3.1 Bayesian Sensory Binding section as well as qualitative, quantitative, and manipulative.

BCI. An abbreviation for "Brain-Computer Interface." The current name of the developing technological field concerned with the direct readout and manipulation of biological neurons. Sometimes extended to refer to more indirect linkages, such as those that attempt to decode electroencephalograms and so on.

cache. To "cache" a result is to store it for later use. For example, your Web browser has a "cache folder" containing files that you have already downloaded from the Internet; when your browser encounters a URL that it's already followed, it can retrieve the file from the cache folder instead of downloading it again.

catastrophic failure of Friendliness. A failure of Friendliness which causes the AI to stop *wanting* to be Friendly; a nonrecoverable error; an error which wipes out, disables, or bypasses the error-recovery mechanisms.

causal goal system. Defined in 3 An Introduction to Goal Systems.

A goal system in which desirability backpropagates along predictive links. If A is desirable, and B is predicted to lead to A, then B will inherit desirability from A,

contingent on the continued desirability of A and the continued expectation that B will lead to A. Since predictions are usually transitive—if C leads to B, and B leads to A, it usually implies that C leads to A—the flow of desirability is also usually transitive.

See supergoal, subgoal, child goal, parent goal, goal-oriented cognition, goal-oriented behavior, goal, and the sections 3 An Introduction to Goal Systems, 5.1 Cleanly Friendly Goal Systems, and 5.2.1 Generic Goal System Functionality.

causal validity semantics. Defined in 5.6.2 Causal Validity Semantics.

An AI that describes verself as the output of a causal process and that regards certain affectors within that process as being more and less valid. See the defining section, 5.6.2 Causal Validity Semantics, for a more complete explanation.

CFAI. An abbreviation used for Creating Friendly AI, this document, a publication of the Singularity Institute for Artificial Intelligence.

child goal. Defined in 3 An Introduction to Goal Systems.

A relation between two goals. If B is a child goal of A, then B derives desirability from A; that is, B is seen as desirable because A is already known to be desirable and B is predicted to lead to A. It does not make sense to speak of a goal as being a "child goal" or "parent goal" in isolation, since B may be a child goal of A while simultaneously being the parent goal of C. See also supergoal, subgoal, parent goal.

cleanly causal goal system. Defined in 5.1 Cleanly Friendly Goal Systems.

A causal goal system in which it is possible to view the goal system as containing only *decisions, supergoals,* and *beliefs,* with all subgoal content being identical with beliefs about which events are predicted to lead to other events. Cleanliness, or an approximation to cleanliness, is a desirable property of Friendship systems; even more important is that a Friendly seed AI should view vis goal system as a *design approximation* to a cleanly causal goal system. I.e., a cleanly causal goal system is normative.

complex functional adaptation. An adaptation composed of multiple, interacting components which work together to perform a function.

Complicated functions do not occur in single mutations —they build up gradually, as simple implementations of the functionality allow for more and more complex improvements to take place. It is impossible to improve on a nonexistent piece of functionality.

A complex functional adaptation necessarily requires more than one mutation to evolve. Successful mutations, or adaptations, occur in response to invariants of

the environment. For a mutation to be successful—for a mutation to be an evolutionary advantage—requires environmental factors that are present with sufficient reliability for a selection pressure to exist. The key concept is that *environment* includes not just *external* environment but internal *genetic* environment. For an adaptation A to occur in response to another adaptation B, adaptation B must be reliably present in the genetic environment, or adaptation A will not be an evolutionary advantage.

Sometimes, when a complex adaptation is just getting started—i.e., there are only two genes comprising the complete adaptation—it is possible for both A and B to float around in a small percentage of the gene pool, especially if A and B are *small* advantages independently but a *large* advantage in combination. Once the percentage rises over a certain threshold level, however, the selection pressures become strong enough that both A and B rapidly become universal within the gene pool. Evolution, while feeding on variance, uses it up. Once both A and B are universal, further adaptations/improvements can then begin to occur in response to the new genetic invariant, the complex adaptation A+B.

The fundamental implication for psychology is that *all complex functional adaptations are panhuman.* Whatever surface differences may result from the variance between humans, all of our ultimate hardware, the basic *design,* is fundamentally identical for all humans. The variance bound up in the Gaussian distribution of intelligence is dwarfed by the vast amount of previously accreted complexity. This is an important thing to keep in mind because our social environment tends to amplify the differences between individual humans and distract us from the differences between humans and chimpanzees.

The idea of a "complex functional adaptation" is one of the deepest concepts of modern-day evolutionary theory. For further explanation, I highly recommend "The Psychological Foundations of Culture" by Cosmides and Tooby.

computational horizon. The scope of a problem. The amount of computing power that needs to be devoted to a task. Outside the "horizon" lie all the facts that are not relevant to the task, all the details that are too fine to be processed with the available computing power, all the consequences that are too hazy or too unimportant to predict, and so on. Deciding where the computational horizon lies often has a major impact on the quality and speed of a cognitive task. See predictive horizon, granularity horizon, and relevance horizon.

computronium. Matter that has been transformed from its natural state into an optimized, maximally efficient computer. (A true Extropian would argue that this *is* matter's "natural state.")

What constitutes "computronium" varies with the level of postulated technology. A rod logic nanocomputer is probably too primitive to qualify as computronium, since large molecular aggregates (hundreds or thousands of atoms) are used as computing elements. A more archetypal computronium would be a three-dimensional cellular automaton which attached computational significance to each individual atom, perhaps with quantum-computing elements included.

More exotic forms of computronium include neutronium, Higgsium, monopolium, or—my personal invention—an interlaced structure of positive-matter and negative-matter monopolium wrapped up in a fractal Van Den Broeck warp. (The total mass is zero, so the whole doesn't undergo gravitational collapse. If paired negative and positive matter can be manufactured in unlimited quantities, the fractal Van Den Broeck warp can continue extending indefinitely and exponentially. Threading the system with wormholes keeps latency down. And the whole thing fits in your pocket.)

configuration space. Suppose you have three numbers, A, B, and C, any one of which can take on any scalar value between 0 and 5. Mathematically, the class of possible states of the system can be viewed as a three-dimensional space, with one dimension for each number. If we're interested in some particular subset of the possible states (for example, the states where A + B + C < 5), this defines a *volume* in *configuration space.*

If one considers a system of a billion quarks, with each quark having three real numbers defining position and three real numbers defining velocity, then the result is a six-billion-dimensional *phase space.* The sets of possible states of the system that can be described as "ice crystal" would constitute a particular volume in this phase space.

conflate. To confuse two conceptually distinct entities by giving them the same name and reasoning about them as if they were the same thing. In the classic philosophical "bait-and-switch," the philosopher argues using one referent of a term, then applies the conclusions to a different referent of the term, while using the same word throughout.

conspecific. Evolutionary-biology jargon for another member of the same species. In our ancestral environment, we weren't just competing with predators, we were competing against our fellow humans. This is illustrated by a famous joke:

Two evolutionary biologists were running frantically from a large, hungry lion. Gasping, the first one said: "This is hopeless. We can't possibly outrun a lion." The second one said: "I don't have to outrun the lion. I just have to outrun you."

Competing against conspecifics is an open-ended process, a Red Queen's Race (Ridley 1994). The best genes of one generation become universal—and therefore merely average—in the next.

Any time you see an enormously exaggerated trait—the tail of a peacock, the antlers of a buck, the intelligence of a human—it's a good guess that two forces are at work: Competition with conspecifics, and sexual selection.

controlled ascent. Defined in 5.7 Developmental Friendliness.

An ethical injunction which states that self-improvement should not outpace programmer guidance, because this increases the probability of catastrophic failure of Friendliness. A consequence of this instruction would be that, if self-improvement suddenly accelerates, it may be necessary to consult the programmers before continuing, or to deliberately slow down for some period of time in order for Friendship to catch up with intelligence.

For very young AIs, "controlled ascent" may be a programmatic feature triggering a save-and-stasis if a metric of self-improvement suddenly accelerates, or if the metric of self-improvement outpaces a metric of programmer guidance.

See 5.8.0.4 Controlled Ascent

Cosmides and Tooby. Leda Cosmides and John Tooby. See The Psychological Foundations of Culture (Tooby and Cosmides 1992).

counterfactual. A what-if scenario deliberately contrary to reality; e.g., "What if I *hadn't* dropped that glass of milk?" See also subjunctive.

crystalline. Defined in Yudkowsky (2001, § 1.1 Seed AI).

Loosely speaking, "crystalline" is the opposite of "rich" or "organic". If vast loads of meaning rest on the shoulders of individual computational tokens, so that a single error can break the system, it's crystalline. "Crystalline" systems are the opposite of "rich" or "organic" error-tolerant systems, such as biological neural networks or seed-AI mindstuff. Error-tolerance leads to the ability to mutate; mutation leads to evolution; evolution leads to rich complexity—networks or mindstuff with lots of tentacles and connections, computational methods with multiple pathways to success.

currently unknown supergoal content. Defined in 5.5.1 External Reference Semantics.

Those portions of the probabilistic supergoals which are unknown or uncertain—see external reference semantics. Subgoals having to do with improv-

ing the supergoals or improving the Friendship structure are child goals of the unknowns in the supergoal content, not super-supergoals or meta-supergoals.

cytoarchitecture. "Cytoarchitecture" refers to the general way neurons connect up in a given lump of neuroanatomy—many-to-many, many-to-one, and so on.

declarative. The distinction between "procedural" and "declarative" knowledge/skill/information is one of the hallowed dichotomies of traditional AI. Although the boundary isn't as sharp as it's usually held to be—especially in seed AI—the distinction is often worth making.

Your knowledge of how to walk is "procedural" and is stored in procedural form. You don't say to yourself: "Right leg, left leg, right leg, left leg." All the knowledge about how to balance and maintain momentum isn't stored as conscious, abstract, *declarative* thought in your frontal lobes; it's stored as unconscious *procedural* thought in your cerebellum and spinal cord. (Neurological stereotyping included for deliberate irony.)

Inside source code, the procedural/declarative distinction is even sharper. A piece of code that turns on the heat when the temperature reaches 72 and turns on the air conditioning when the temperature reaches 74, where 72 and 74 are hard-coded constants, has *procedurally* stored the "correct" temperature of 73. The number "73" may not even appear in the program.

A piece of code that looks up the "target temperature" (which happens to be 73) and twiddles heat or A/C to maintain that temperature has *declaratively* stored the number 73, but has still *procedurally* stored the method for maintaining a temperature. It will be easy for the programmer—or for internally generated programs—to refer to, and modify, the "target temperature". However, the program still doesn't necessarily know how it's maintaining the temperature. It may not be able to predict that the heat will go on if the temperature reaches 72. It may not even know that there's such a thing as "heat" or "air conditioning". All that knowledge is stored in procedural form—as code which maintains a temperature.

In general, procedural data is data that's opaque to the program, and declarative data is data that the program can focus on and reason about and modify. Seed AIs blur the boundary by analyzing their own source code, but this doesn't change the basic programmatic truth that declarative=good and procedural=bad.

Deep Blue. The chess-playing device that finally beat the human champion, Kasparov. A great, glorified search tree that beat Kasparov essentially through brute force, examining two billion moves per second. Built by IBM.

desirability. Defined in 5.1 Cleanly Friendly Goal Systems.

Intuitively, a quantitative measure of how much a mind wants to achieve a particular state, irrespective of whether that state is desired for itself or as a means to an end. "Undesirability" is how much a mind wants to avoid a state.

Formally, "desirability" is a quantity that springs from supergoals, flows backward along predictive links to subgoals, and eventually collects in decisions, where the action with the highest "desirability" is always taken. In a causal goal system, "desirability" might be better translated as *leads-to-supergoal-ness*. Leads-to-supergoal-ness is a quantity that usually but not always flows backward along predictive links; i.e, if C is predicted to lead to B, and B is predicted to lead to supergoal A, then C usually (but not always) inherits leads-to-supergoal-ness or "desirability" from B.

Devil's Contract Problem. Defined in 4.7 Interlude: Beyond the Adversarial Attitude.

A failure of Friendliness scenario in which the AI follows the letter but not the spirit of a literally worded order, in the tradition of much older tales about accepting wishes from a djinn, negotiating with the fairy folk, and signing contracts with the Devil. This FoF scenario is raised *far* too often given its relative severity and probability.

In the *diabolic* variant, the misinterpretation is malicious. The entity being commanded has its own wishes and is resentful of being ordered about; the entity is constrained to obey the letter of the text, but can choose among possible interpretations to suit its own wishes. The human who wishes for renewed youth is reverted to infancy, the human who asks for longevity is transformed into a Galapagous tortoise, and the human who signs a contract for life everlasting spends eternity toiling in the pits of hell.

In the *golemic* variant, the misinterpretation is due to lack of understanding. A golem is a made creature which follows the literal instructions of its creator. In some stories the golem is resentful of its labors, but in other stories the golem misinterprets the instructions through a mechanical lack of understanding—digging ditches ten miles long, or polishing dishes until they become as thin as paper. (I stole those two examples from Feet of Clay by Terry Pratchett [1996] rather than the traditional folklore, but they are fine examples nonetheless.)

e.g. *Exempli gratia;* Latin, "for example."

ethical injunction. Defined in 5.2 Generic Goal Systems.

An injunction that is never supposed to be violated—has no preconditions for violation —because the probability of mistaken violation is greater than the probability of correct violation, generally because the injunction compensates for a

known bias in cognition (hopefully a *humans-only* issue). (Actually, the human version gets a bit more complicated—issues of honor and reputation; psychological strength derived from context-insensitivity; and so on.)

In Bayesian terms, given the fact that violating the injunction appears to be fully justified, it is more likely that a cognitive error has occurred than that the action actually is justified. Or rather, considered abstractly and in advance of a specific instance, the total payoff from avoiding even apparently valid violations is expected to be positive.

Since AIs—unlike humans—are in more danger from stupidity than observer bias, and there is no real equivalent to an "irrevocable commitment," the primary utility of ethical injunctions in AIs is that it enables the *current* AI to cooperate with the programmer in safeguarding against *future* violations of an ethical injunction. See 5.3.3.1 Cooperative Safeguards.

Eurisco. Eurisko was the first truly self-enhancing AI, created by Douglas B. Lenat. Eurisko's mindstuff—in fact, most of the AI—was composed of heuristics. Heuristics could modify heuristics, including the heuristics which modified heuristics.

I've never been able to find a copy of Eurisko's source code, but, by grace of Jakob Mentor, I have obtained a copy of Lenat's original papers. It turns out that Eurisko's "heuristics" were arbitrary pieces of LISP code. Eurisko could modify heuristics because it possessed "heuristics" which acted by splicing, modifying, or composing—in short, mutating—pieces of LISP code. Many times this would result in a new "heuristic" which caused a LISP exception, but Eurisko would simply discard the failed heuristic and continue. In a sense, Eurisko was the first attempt at a seed AI—although it was far from truly self-swallowing, possessed no general intelligence, and was created from crystalline components.

Engines of Creation (by K. Eric Drexler) contains some discussion of Eurisko's accomplishments.

external reference semantics. Defined in 5.5.1 External Reference Semantics.

The behaviors and mindset associated with the idea that supergoals can be "wrong" or "incomplete"—that the current supergoals are an approximation to an ideal, rather than being correct by definition.. Under a system with external reference semantics, supergoal content takes the form of probabilistic hypotheses about an external referent. That is, the supergoal content consists of *hypotheses about Friendliness* rather than a *definition of Friendliness.*

The most important and fundamental use of ERS is to protect the ability of the programmers to cooperatively modify and improve the goal system. Letting

the programmers make alterations to the supergoals is not a subgoal of a "let the programmers make alterations" meta-supergoal; it is a child goal of the unknown portion of the supergoal content, or the fact that statements about the supergoal are probabilistic. Technically, a Friendly AI doesn't have supergoal "content"; the Friendly AI has an *external referent* labeled "Friendliness," and a series of probabilistic assertions about that external referent, derived from sensory information such as programmer affirmations.

Explanations of how to "ground" the external reference include 5.6.1 Shaper/Anchor Semantics and 5.6.2 Causal Validity Semantics. The simplest (but ungrounded) form of external reference is an assertion that some external source contains the ideal, or the best available approximation to the ideal, and that sensory data from that source is Bayesian information about what the internal content should be.

The idea is that programmers know about "Friendliness"—have a connection to the external referent—and that statements by the programmers are thus sensory information about external referent "Friendliness."

The term "external reference semantics" derives from the way that many of the behaviors associated with probabilistic supergoals are those associated with probabilistic subgoals—in general, with an uncertain view of external reality. If your current image of reality is an approximation to real reality, or if your current subgoals are dependent on an uncertain image of reality, then blindly defending the current subgoals is undesirable—it could interfere with your ability to create new, better subgoals. Similarly, under ERS, blindly defending hypotheses about Friendliness is undesirable; it could interfere with the ability to achieve the actual Friendliness referent.

extraneous cause. Defined in 5.6.2 Causal Validity Semantics.

A concept from 5.6.2 Causal Validity Semantics. When an AI (or a normative human) creates a plan, ve starts with a mental image of the intended results and designs a series of actions which, applied to the world, should yield a chain of causality that ends in the desired result. If an *extraneous cause* comes along and disrupts the chain of causality, the AI must take further actions to preserve the original pattern; the pattern that would have resulted if not for the extraneous cause. The AI must re-run the chain of causality without the error to find the valid result—a "causal rewrite." Implicit in the idea of an extraneous cause is the idea of a *valid* cause, an idea which forms the foundation of causal validity semantics.

failure of Friendliness. A mistake causing an otherwise Friendly AI to take one or more unFriendly actions, perhaps due to an error in supergoal content or a misapprehen-

sion about external reality. The error may or may not be recoverable. See catastrophic failure of Friendliness.

flight recorder. Defined in 5.3 Seed AI Goal Systems.

For a project with very high-level funding: A complete record of all inputs, sensory data (with precise timing), Internet feeds, keystrokes (with precise timing), source-code checkins, randseeds, and hardware configuration, such that sufficient data exists to reconstruct the AI's program state at any instant in time. Written to WORM (write-once-read-many) memory and stored securely. Using a full flight recorder, it is possible to verify that no radiation bitflips or unrecorded inputs occurred from the AI's moment of conception to vis present-day existence, albeit at great cost in computing time.

For a project with less extravagant funding, a "poor man's flight recorder" is an autonomous subsystem of the AI that writes the AI's current stream of consciousness, in human-readable format, to non-AI-tamperable storage, plus whatever daily or weekly backup snapshots are stored.

FoF. Abbreviation for "failure of Friendliness."

Friendliness. Intuitively: The set of actions, behaviors, and outcomes that a human would view as benevolent, rather than malevolent; nice, rather than malicious; friendly, rather than unfriendly; good, rather than evil. An AI that does what you ask ver to, as long as it doesn't hurt anyone else, or as long as it's a request to alter your own matter/space/property; an AI which doesn't cause involuntary pain, death, alteration, or violation of personal environment.

Friendly AI. (1) The field of study concerned with the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals. The term "Friendly AI" was chosen not to imply a particular *internal* solution, such as duplicating the human friendship instincts, but rather to embrace *any* set of external behaviors that a human would call "friendly." In this sense, "Friendly AI" can be used as an umbrella term for multiple design methodologies. Usage: "The field of Friendly AI."

(2) An AI which was designed to be Friendly. Within the context of *Creating Friendly AI,* an AI having the architectural features and content described in this document. Usage: "A Friendly AI would have probabilistic supergoals."

(3) An AI which is currently Friendly. See Friendliness. Usage: "The first AI to undergo a hard takeoff had better be a Friendly AI."

Friendship. The systems, subsystems, goal system content, system architecture, and other design features constituting the implementation of Friendliness.

Friendship architecture. Defined in 2 Challenges of Friendly AI.

The challenges of Friendship acquisition and Friendship structure, as distinguished from Friendship content. The part of the problem that is solved by creating specific cognitive processes, rather than by adding specific knowledge or other cognitive content. The bounded amount of complexity that is infused by design and forethought, rather than the open-ended amount of complexity that is accumulated through experience. See 2.4 Content, Acquisition, and Structure.

Friendship acquisition. Defined in 2 Challenges of Friendly AI.

The second-order problem of Friendly AI; building a Friendly AI that can *learn* Friendship content. The cognition used to verify, modify, improve, and contradict Friendship content. (Note that "Friendship content" thus has the connotation of something that can be verified, modified, improved and contradicted without that posing an unusual problem.) See 2.4 Content, Acquisition, and Structure and Friendship structure.

Friendship content. Defined in 2 Challenges of Friendly AI.

The zeroth-order and first-order problems of Friendly AI; correct decisions and the cognitive complexity used to make correct decisions. The complex of beliefs, memories, imagery, and concepts that is used to actually make decisions. Specific subgoal content, supergoal content, shaper content, and so on. See 2.4 Content, Acquisition, and Structure; see Friendship acquisition and Friendship structure.

Friendship structure. Defined in 2 Challenges of Friendly AI.

The third-order problem of building a Friendly AI that *wants* to learn Friendliness (engage in Friendship acquisition of Friendship content). The structural problem that is unique to Friendly AI. The challenge of building a funnel through which a certain kind of complexity can be poured into the AI, such that the AI *sees that pouring as desirable* at every point along the way. The challenge of creating a bounded amount of Friendship complexity that can grow to handle open-ended philosophical problems. See 2.4 Content, Acquisition, and Structure.

gender-neutral pronouns. Ve, vis, ver, verself. I was forced to start using gender-neutral pronouns when referring to intelligent AIs, since to use "he" or "she" would imply cognitive hardware that such an AI would very specifically *not* have.

I realize that these pronouns strike people as annoying the first time around. I'm sorry for that, and I truly regret having to annoy my readers, but "it" is simply

inadequate to refer to AIs. Not only is "it" used as a pronoun for inanimate matter, but "it" is also a general anaphor, like "this" or "that". "It" can refer to anything at all in a sentence, not just the AI, so complex sentences—especially ones that use "it" for other purposes—become impossible to parse syntactically. Sometimes a sentence can be rewritten so that no pronoun is necessary, but for sentences with multiple pronoun references, this rapidly becomes either impossible, or too tangled. I would rather use unusual words than tangled syntax. At least "ve" gets easier to parse with time.

At one point I was using "ve" to refer to a human of indefinite gender, but I have since realized that this is just as inaccurate as referring to an AI as "he" or "she". I now keep a coin near my computer that I flip to decide whether a human is "he" or "she". (No, I haven't fallen into the bottomless pit of political correctness. Everyone has the right to use whatever language they like, and I can flip a coin if I want to. Your right to use "he" implies my right to flip a coin. Right?)

GEB. Gödel, Escher, Bach: An Eternal Golden Braid by Douglas R. Hofstadter. This book is *mandatory* reading for all members of the human species.

giga. 10^9. One billion; a thousand "mega."

GISAI. An abbreviation used for *General Intelligence and Seed AI* (Yudkowsky 2001), a publication of the Singularity Institute for Artificial Intelligence. Located at http://singularity.org/files/GISAI.html

goal. Defined in 3 An Introduction to Goal Systems.

A state of the world marked as desirable (or undesirable), either because the world-state is inherently desirable under the supergoal content, or because the world-state is necessary as the prerequisite of some other desirable state. An entity exhibiting goal-oriented behavior will act so as to reach a "goal" state; a mind capable of goal-oriented cognition will take actions which ve *predicts will lead to* the goal state. Usually, "goal" implies a mind capable of goal-oriented cognition.

goal-oriented behavior. Defined in 3 An Introduction to Goal Systems.

Behavior that leads the world towards a particular state. Behavior which *appears* deliberate, centered around a goal or desire.

The perception of goal-oriented behavior comes from observing multiple actions that coherently steer the world towards a goal; or singular actions which are uniquely suited to promoting a plausible goal-state and too improbable to have arisen by chance; or the use of different actions in different contexts to achieve a single goal on multiple occasions.

A more technical definition may be offered as follows: An organism or device exhibits "goal-oriented" behavior when, across multiple starting locations in a volume of the configuration space of the local world, the behavior produced by the organism or device differs depending on the starting location, such as to steer the final result of the local world toward a more compact volume of configuration space. (Note that in a thermodynamic Universe, the organism/device must produce waste entropy in order to steer the local world toward a more compact ending state. Generally, the perception of goal-oriented behavior arises when *high-level* chaos is converted into *high-level* order, without the observer tracking lower levels of description. For organisms/devices and observers in a low-entropy Universe, the organism/device should probably be conceived of as affecting the configuration space of a high-level description abstracted from the local world.) See also goal-oriented cognition and 3.1 Interlude: The Story of a Blob.

goal-oriented cognition. Defined in 3 An Introduction to Goal Systems.

Goal-oriented cognition describes a mind which possesses a mental image of the "desired" state of the world, and a mental image of the actual state of the world, and which chooses actions such that the projected future of world-plus-action leads to the desired outcome state. See goal-oriented behavior and manipulative.

granularity horizon. The fineness of the detail that needs to be modeled. How much reductionism needs to be applied to capture all the relevant details. The tradeoff between expenditure of computing power, and the benefits to be gained from finer modeling.

Every time your eye moves, the amount of processing power being devoted to each part of the visual field changes dramatically. ("The cortical magnification factor in primates is approximately inversely linear, at least for the central twenty degrees of field." "It has been estimated that a constant-resolution version of visual cortex, were it to retain the full human visual field and maximal human visual resolution, would require roughly 10^4 as many cells as our actual cortex (and would weigh, by inference, roughly 15,000 pounds)." MITECS, "Computational Neuroanatomy".) And yet we can watch an object rotating, so that different parts move all over the visual cortex, and it doesn't appear to distort.

Every time you change scale or the level of detail at the modality level of representation, the data may wind up going into essentially a different format—at least, from the perspective of someone trying to detect identity by bitwise comparison. Even adding or subtracting pieces of the puzzle, without changing scale, can be a problem if the AI has painstakingly built up a perceptual tower that doesn't take well to tampering with the foundations. Cognitive methods need to be able to

take these kinds of random pushes and shoves. "Error-tolerance" isn't just important because of actual *errors*, but because all kinds of little flaws naturally build up in a cognitive task, as the result of cognition.

See computational horizon.

grok. Hacker slang for "understand" or "fully, completely understand" or "understand in all its forms and ramifications" or "embrace." The connotation is that the understanding is not just abstract and removed, but detailed enough to be immediate and fully usable. Originates in Robert Heinlein's *Stranger in a Strange Land*.

hard takeoff. The Singularity scenario in which a mind makes the transition from prehuman or human-equivalent intelligence to strong transhumanity or superintelligence over the course of days or hours. See Yudkowsky (2001, § 1.1 Seed AI), Yudkowsky (2001, § What is Seed AI?), or http://intelligence.org/ie-faq/.

hertz. A measure of frequency, equal to 1 cycle per second. A neuron that fires 200 times per second is operating at 200 Hz. CPU clock speeds are currently measured in megahertz (MHz) or gigahertz (GHz).

heuristic. I use the term to refer to any piece of knowledge which provides a rule of thumb—anything from "Don't rest your hand on a hot stove" to "Try to control the center of the chessboard".

Some other definitions: Douglas Lenat once wrote an AI called Eurisko in which the mindstuff —in fact, practically the entire AI—was composed of "heuristics" which could modify other heuristics, including the heuristics doing the modifying. For example, "investigate extreme cases" was modified by a heuristic to yield "investigate cases *close* to extremes." (Douglas Lenat went on to state that"Heuristics are compiled hindsight; they are judgmental rules which, if only we'd had them earlier, would have enabled us to reach our present state of achievement more rapidly.") In classical AI, a"heuristic" is usually a function used to prune search trees by indicating branches which are likely or unlikely to be desirable.

holism. The attitude that the whole is greater than the sum of the parts. To take the holistic view is to look upward, focus on the high-level properties. See reductionism and reductholism. See also Gödel, Escher, Bach: An Eternal Golden Braid, particularly the dialogues "Prelude" and "Ant Fugue".

"No one in his right mind could deny holism." —GEB

human-equivalence. A threshold level of ability in some domain which, if achieved, allows the AI to understand human concepts and do work as good as a human in that domain. A human-equivalent AI is one that has passed the threshold level of

general competence in some sense. An ambiguous and arguable definition; "transhumanity," especially strong transhumanity, is more blatant.

iff. "Iff" is shorthand for "if-and-only-if."

inclusive reproductive fitness. A measure of adaptivity or fitness which takes into account effects on close relatives. J.B.S. Haldane was once asked if he would risk death by drowning to save his own brother. He famously replied: "No, but I would to save two brothers or eight cousins." (A sibling shares, on average, half your genome; a first cousin shares an eighth of your genome.)

infrahuman AI. An AI of below-human ability and intelligence. "Prehuman" is usually referred to use to an infantlike or tool-level AI, so that "infrahuman" usually refers to a fairly mature AI, capable of general cognition, which is still not in the vicinity of human intelligence. See prehuman, near-human, human-equivalent, transhuman, and superintelligent.

injunction. Defined in 5.2 Generic Goal Systems.

A planning heuristic which has at least partial nonlocal support, or a planning heuristic which exhibits a great deal of context-independence. The archetypal case would be a heuristic that is supposed to be applied even when the straightforward interpretation of the world-model suggests otherwise, generally (in AIs) due to unknown unknowns or (in humans) to compensate for framing effects or (for both) to save computing power. See ethical injunction.

instantiate. Loosely, program A "instantiates" program B if it can perfectly simulate program B. An "instantiation" of a program is a running copy of that program.

This issue actually gets waaay more complicated, but I'm not going to inflict that on you now. Maybe later. Nobody has ever come up with a mathematical definition of "instantiation", but it's a useful concept.

intelligence. Defined in Yudkowsky (2001, § 2.1 World-model).

What is intelligence? In the case of humans, intelligence *is* a brain with around 40 billion neurons, and 104 cytoarchitecturally distinct areas in the cerebral cortex alone. What intelligence *is* is the subject of this whole web page.

The *cause* of intelligence can be more succinctly described: Evolution is the cause of intelligence, and intelligence is an evolutionary advantage because it enables us to model, predict, and manipulate reality. Or rather, it enables us to model, predict, and manipulate *regularities* in reality.

Kasparov. Gary Kasparov, the human who finally lost the world chess championship to Deep Blue.

kilo. 10^3. One thousand.

latency. *Latency* describes delays, specifically delays introduced by *communication* rather than local processing, and *irreducible* delays rather than delays caused by sending a large amount of data. Most commonly used in discussion of computer networks and hardware systems. The *latency* on a motherboard is, for example, the time it takes a message from the CPU to reach a video card. The *latency* between nodes of a network is the time it takes for a message from node A to reach node B. For a fine explanation of why "latency" is entirely distinct from "bandwidth"—adding an identical second channel doubles bandwidth but does not affect latency at all—see "It's the Latency, Stupid."

Note that our Universe specifies that the minimum latency between two nodes will be at least one second for every 186,000 miles. The latency between two nodes 300 microns apart must be at least one picosecond.

Anders Sandberg, in "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains," suggests the measure *S* of the "diameter" of a single mind, where *S* is the ratio of the clock speed to the latency between computing elements. (*S* = distance / (clock speed * message speed)). Anders goes on to note that the human brain has $S \sim 1^{96}$). An oft-voiced conjecture is that the subjective "feel" of having a single, unified mind may require $S <= 1$. (As far as I know, this conjecture is only applied to superintelligences, and nobody has suggested that *S* played a role in shaping human neurology.)

Law of Pragmatism. Defined in Yudkowsky (2001, § 1.2 Thinking About AI).

Any form of cognition which can be mathematically formalized, or which has a provably correct implementation, is too simple to contribute materially to intelligence.

life. So, you want to know the meaning of life, eh?

When capitalized, "Life" usually refers to Conway's Game of Life, a two-dimensional cellular automaton. Cells are laid out in a square grid, and cells can either be alive or dead. Each cell is affected only by the eight cells around it. With each tick, these rules are applied to each cell:

1: A cell with fewer than two living partners becomes or remains dead.

2: A cell with two living parters maintains its current state.

3: A cell with three living partners becomes or remains alive.

96. 200 Hz neurons, 100 m/s axons, and a 0.1m diameter

4: A cell with four or more living partners becomes or remains dead.

These rules are enough to generate almost endless variance.

As for the age-old controversy among biologists about how to define "life", I suggest the following: "Life is anything designed primarily by evolution, plus anything that counts as a person." Note that this definition includes mules, excludes computer viruses, includes biological viruses, and (by special clause) includes designed minds smart enough to count as people.

LISP tokens. LISP is a programming language, the traditional language of AI.

"LISP" stands for "List Processor". In LISP, everything is made from lists, including the code. For example, a piece of code that adds 2 and 2 would be `(plus 2 2)`. This code is a list composed of three tokens: `plus`, `2`, and `2`. If `"num1"` was a LISP token that contained the value of `2`, the code could be written `(plus num1 2)` and would return `4`.

When I say that classical AI is built from suggestively-named LISP tokens, I mean that the classical AI contains a data structure reading `((is(food hamburger)) (is(eater human)) (can-eat (eater food)))`; the classical AI then deduces that a `"human"` can `"eat"` a `"hamburger"`, and this is supposed to be actual knowledge about hamburgers and eating. What the AI really knows is that a `G0122` can `H8911` a `G8733`. Drew McDermott pointed this out in a famous article called "Artificial Intelligence Meets Natural Stupidity".

(LISP does have one real property which is important to AI, however; the code and the data structures follow the same format, making LISP the single premier language for self-modifying code. A true AI would probably read C++ as easily as LISP, since the amount of complexity needed to parse code is comparatively trivial relative to the amount of cognitive complexity needed to *understand* code. Even so, using a language well-suited to self-modification may simplify the initial stages of the AI where self-improvement is mostly blind. Since LISP is getting ancient as programming languages go, and linked lists are awkward by today's standards, I've proposed a replacement for LISP called "Flare", which (among many other improvements) would use XML instead of linked lists. Even so, of course, putting faith in the token level of Flare would be no better than putting faith in the token level of LISP. At most, Flare might be well-suited to programming sensory modalities and mindstuff. It would be nice to have such a language, since none of the existing languages are really suited to AI, but it's more likely that we'll just hack something up out of Python—during the initial stages, at least. For more about Flare, see the obsolete document The Plan to Singularity.)

mature nanotechnology. See nanotechnology. Nanosystems (Drexler 1992) describes, for example, acoustic nanocomputers using "rod logics" composed of diamondoid rods with hundreds or thousands of atoms each, moving at the speed of sound in diamond, providing 10^{21} ops/sec from a one-kilogram nanocomputer.

This level of nanotechnology is best thought of as representing the nanotechnological equivalent of vacuum tubes. "Mature nanotechnology" is nanotechnology that bears the same resemblance to rod logics as, say, a modern ultra-optimized VLSI chip bears to ENIAC. This would cover electronic nanocomputers verging on computronium, extremely small (~100 nanometer or ~10 nanometer, rather than ~1 micron) nanobots capable of swarming to perform complex manufacturing activities (including reproduction) in a natural non-vat environment, ultra-optimized molecular specifications, error-tolerant nanomachinery, and so on.

mega. 10^6. One million; a thousand "kilo."

meme. An idea considered as an organism, or focus of selection pressures. Ideas reproduce by being linguistically transmitted from one human to another; they mutate when individuals make errors in the retelling. If most of the content of an idea is the result of adaptive transmission errors, or if the idea's content is best explained by reference to evolutionary phenomena rather than the decisions of individuals, then that idea is a meme. Colloquially, ideas are also referred to as "memes" when they are being thought about as memes rather than thoughts—i.e., when the important characteristics of an idea is its reproductive fitness or cultural spread, rather than its content. Although the term is usually derogatory, memes can sometimes be positive forces as well—for example, when a meme relies on positive emotions or truthful thinking to compete. Lies and hate are the easier and faster Dark Side, though, so the term is usually derogatory.

Mirror. An as-yet-unimplemented idea for training humans to think more intelligently by finding localizable hardware substrate for such cognitive adaptations as hatred and rationalization, then training humans to deliberately avoid the use of those abilities by asking them to philosophize in an fMRI machine. A possible futuristic method for training Friendly AI programmers, if the basic method is feasible and the Earth sticks around that long.

MITECS. The MIT Encyclopedia of the Cognitive Sciences. (Wilson and Keil 1999). A truly excellent book containing 471 short articles about topics in the cognitive sciences. See also my review in the Bookshelf.

moral deixis. Defined in 5.3 Seed AI Goal Systems.

MITECS, "Context and Point of View":

> ". . . in *The CIA agent knows that John thinks that a KGB agent lives across the street* . . . the underlined phrase can be evaluated from the speaker's, the CIA agent's, or John's point of view.
>
> "Three major kinds of deixis are usually distinguished: (1) *person deixis,* where the context provides one or more participants of the conversation . . . (2) *spatial deixis,* where the context provides a location or a direction, especially as a reference point for spatial orientation . . . (3) *temporal deixis,* with the context contributing a specific time . . ."

Deixis is the use and communication of statements which contain speaker variables. *Moral deixis* refers to the (human) use, communication, and expectation of moral statements which contain speaker variables.

If John Doe says to Sally Smith "My philosophy is: 'Look out for John Doe.' ", Sally Smith will hear "Your philosophy should be: 'Look out for Sally Smith.' ", not "Your philosophy should be: 'Look out for John Doe.' " What has been communicated is "Look out for *[speaker],"* a moral statement whose specific content varies among each listener due to moral deixis. Our instinctive expectation of moral deixis is very strong; also, we take for granted (a) that anyone enunciating a moral principle is trying to get others to adopt it, and (b) that any moral principle thus enunciated will be phrased so as to appeal to others. Thus, hearing (from John Doe) 'Look out for John Doe', we automatically assume that 'John Doe' is the result of speaker deixis—a particular value of a variable—and that the moral principle actually being enunciated is "Look out for *[speaker].*" A third party, Pat Fanatic, may be able to say:"My philosophy is: Worship the great leader John Doe", and this will be heard by listeners as"Look out for [John Doe]," but it is conversationally impossible for John Doe himself to communicate this information.

A human, trying to achieve unity of will with a Friendly AI—for purposes of prediction or creation—must *simultaneously* suppress her adversarial attitude (speaker deixis in personal supergoals) *and* suppress her instinctive expectation that the AI will make observer-biased decisions. This is called "eliminating moral deixis," one of the greater prerequisites —or challenges—of the art of Friendly AI.

nanocomputer. A computer built using nanotechnology (manufacturing to molecular specifications). A lower bound on nanocomputing speeds has been set by calculating the speed of an acoustic computer using "rod logics" and messages that travel at the speed of sound; a one-kilogram rod logic, occupying one cubic centimeter, can contain 10^{12} CPUs each operating at 1000 MIPS for a total of a thousand billion billion operations per second. See rod logic.

Note that rod logics are the nanotech equivalent of vacuum tubes (circa 1945), or rather, Babbage's Analytical Engine (circa 1830). Electronic nanocomputers would be *substantially* faster. We use the "rod logic" numbers because they're easy to analyze, and because 10^{21} operations per second are sufficient for most applications.

nanotechnology. Developing molecular nanotechnology would mean the ability to synthesize arbitrary objects to atomic-level specifications. For an introduction, see Engines of Creation (Drexler (1986), now online). For a technical visualization of *lower* (not upper) limits on the potential of nanotechnology, see the book Nanosystems (Drexler 1992). These lower limits—the nanotechnological equivalent of vacuum tubes—include a one-kilogram computer, running on 100 kW of power, consisting of 10^{12} CPUs running at 10^9 ops/sec, for a total of 10^{21} ops/sec. By comparision, the human brain is composed of approximately 100 billion neurons and 100 trillion synapses, firing 200 times per second, for a total of somewhere around 10^{17} ops/sec. See rod logics.

Nanosystems also describes molecular manufacturing systems capable of creating copies of themselves in less than an hour. This implies a certain amount of destructive potential. An exponentially replicating assembler could reduce the biosphere to dust in a matter of days. (For an overly optimistic treatment of the problem, see "Some Limits to Global Ecophagy" by Robert Freitas Jr. [2000].) Accidental out-of-control replication is fairly easy to prevent, given a few simple precautions; we should be more worried about military-grade nanotechnology and deliberately developed weapons. Given our human propensity to make things that go bang—and use them on each other—it would probably be a good idea to develop AI *before* nanotechnology. Nanotechnology is the "deadline" for AI.

In the long run, there are only two kinds of technology: There are technologies that make it easier to destroy the world, and there are technologies that make it possible to go beyond the human. Whether we possess even a *chance* of survival is a question of which gets developed first. Developing a transhuman AI does involve certain risks, but it's better than the alternative—success in Friendly AI improves our chances of dealing with nanotechnology much more than success in nanotechnology would improve our chance of creating Friendly AI.

See 6 Policy Implications.

near-human AI. An AI roughly in the vicinity of human intelligence, only slightly above or slightly below, or perhaps with some transhuman and some infrahuman abilities. Capable of interacting with humans as game-theoretical equals. Under a hard takeoff scenario, near-human AIs would exist very briefly, if at all.

nondestructive brain scan. As a theoretical construct, the idea that a Friendly superintelligence armed with mature nanotechnology has the theoretical capability to perform a complete (nondestructive) scan of all the neurons and synapses in the programmer's mind and thus get a *complete* readout of mind-state, enough to settle *absolutely* any lingering questions about what the programmer really wanted and why. A nondestructive brain scan shouldn't *really* be necessary—any strongly transhuman AI should be quite capable of understanding a human mind. However, the thought experiment does show that any unresolved questions of causal validity and shaper/anchor semantics should not linger into the "nanotech-capable superintelligence" stage.

nonmalicious mistake. Defined in 5.2 Generic Goal Systems.

A failure of Friendliness which is clearly visible as a failure of Friendliness if envisioned by the Friendly AI; a mistake made due to inadequate predictive horizons or factual misapprehensions rather than a failure in the goal system. A straightforward failure by the goal system to achieve its own supergoals, rather than mutation of supergoal content or misunderstanding of supergoal definitions and so on.

normative. A term used by cognitive scientists to contrast the ideal or "normative" forms of cognition with the psychologically realistic cognition occurring in humans. For an example, see the MITECS article quoted at the beginning of 5.2.4.1 Anthropomorphic Injunctions.

objective. Existing outside the mind, in external reality.

Truly objective facts, things that are *really real,* should have binary, all-or-nothing existence or nonexistence; the truth or falsity of a statement entirely about objective facts should be either 100% or 0%. The *only* objects in this class are quarks, other fundamental particles, possibly spacetime, possibly the laws of physics, maybe mathematical facts,[97] and in some theories qualia. Objective facts inhabit the ultimate lowest level of external reality—the token level of our ontology.

In the nonabsolute sense, "objective" is often used to indicate a statement which has a fairly direct correspondence (sensory binding) to reality; a statement which makes reference to external reality rather than internal cognitive structures; which tends toward observer-independence rather than the reverse; and whose truth or

97. It is possible that mathematics is objective. It is also possible that the mathematical laws we observe are merely derivative patterns; higher-level behaviors exhibited by the true laws of physics.

falsity is more dependent on variables in external reality than on variables in the mind of the observer.

Opposed to subjective.

objective morality. Defined in 5.6.2 Causal Validity Semantics.

A set of entirely factual statements which will convince any general intelligence, even an entirely passive system containing no differential desirabilities, to pursue a specific (or perhaps specifiable) supergoal. Under another variation, a set of entirely factual statements which will convince any general intelligence, even one that *already has directly opposing supergoals,* to pursue some specific supergoal. The tail-end recursion of the "rule of derivative validity"; the Meaning of Life. Probably nonexistent—in fact, provably nonexistent, but the same proof of nonexistence applies to the First Cause, so we probably can't be entirely sure. See the defining section, 5.6.2.6 Objective Morality, Moral Relativism, and Renormalization. (Those of you who have read my past works should note that renormalization, not objective morality, forms the proposed ultimate basis of Friendly AI.)

observer-biased. A perceived quantity that tends to assume values whose perception will benefit the perceiver. Evolved organisms, particularly imperfectly deceptive social organisms, tend to develop observer biases.

Things that are observer-biased (unless you have extreme levels of mental discipline or you're a Friendly AI):

- The correctness of your political views.
- The likelihood that you are to blame for any given negative outcome.
- Your own trustworthiness, relative to the trustworthiness of others.
- Your likelihood of success if placed in charge of a social endeavor.

See observer-dependent, observer-independent, observer-centered, and Yudkowsky (2001, § Interlude: The Consensus and the Veil of Maya).

observer-centered. Cognitive content with significant use of speaker deixis; cognitive processes which make significant use of the point of view.

Observer centering is a stronger form of "observer dependence"—a perception can vary from observer to observer without necessarily being centered on each observer. A dozen people walking around a building, each getting a different visual view of that building, are receiving observer-dependent visual information. A dozen people watching a sunset on the beach, each seeing a trail of sparkles on

the water that leads directly to him, are receiving observer-centered visual information. *Goals* in evolved organisms tend to be observer-*centered*; beliefs tend to be observer-*biased.* See also observer bias and moral deixis.

observer-dependent. A perceived quantity which varies depending on who does the perceiving. Subjective, as opposed to objective.

Things that are observer-dependent:

- The speed of a passing train.[98]
- The length of a ruler.[99]

Things that *may* be observer-dependent:

- Whether vanilla or chocolate tastes better.

See observer-independent, observer-biased, and Yudkowsky (2001, § Interlude: The Consensus and the Veil of Maya).

observer-independent. A perceived or physical quantity to which all rational observers will assign a single value. Objective, as opposed to subjective.

Things that are observer-independent:

- The existence of a quark.
- The sum of 2 + 2. (Maybe!)
- The speed of light in a vacuum.[100]

See observer-dependent, observer-biased and Yudkowsky (2001, § Interlude: The Consensus and the Veil of Maya).

ontology. The basic level of reality. Our ontology involves quarks, spacetime, and probability amplitudes. The ontology of a Life game consists of dead cells, live cells, and the cellular-automaton rules. The ontology of a Turing machine is the state transition diagram, the read/write head, and an infinitely long tape with ones and zeroes written on it.

orthogonal. A mathematical term; in geometry, it means perpendicular. Colloquially, two variables that can change independently of each other; not necessarily mutually irrelevant, but decoupled.

98. How fast were *you* going when you made the measurement?

99. Another Special-Relativity thing. If you run past the ruler at 75% of the speed of light, measuring the ruler will show that it's around half the length you would otherwise expect.

100. At least, the *local* speed of light. I'm not getting into General Relativity.

oversolve. Defined in 5.3 Seed AI Goal Systems. A philosophy under which, once you solve a problem, you go on to solve the problem with half your brain tied behind your back. See wisdom tournaments. If the AI can make Friendly decisions while running on unreliable hardware using deliberately messed-up software with half the heuristics disabled, a number of pseudo-human biases tossed in, and all the programmer-created ethical injunctions switched off—not just make the right decisions, but make them cleanly, simply, and with plenty of margin for error, so that the messed-up AI was never remotely within distance, not of catastrophic failure, but of a single moral compromise—then presumably the *real* AI, running at full power and full intelligence, has a "strength of Friendly personality" almost beyond human comprehension. That's "oversolve."

parent goal. Defined in 3 An Introduction to Goal Systems.

A relation between two goals. If B is a parent goal of C, then C derives desirability from B; that is, C is seen as desirable because B is already known to be desirable and C is predicted to result in B. It does not make sense to speak of a goal as being a "parent goal" or "child goal" in isolation, since B may be the parent goal of C while simultaneously being a child goal of A. See also supergoal, subgoal, child goal.

past light cone. The set of all events in causal contact with a given spacetime point. The past light cone is the space of all events from which a ray of light could have reached the current event. The future light cone is the space of all events that can be reached by a ray of light from the current event. Any event occurring outside our past light cone can have no causal impact on this moment.

predictive horizon. How far into the future the consequences of an event need to be projected. The amount of computing power devoted to projecting the consequences of an action. In Friendly AI, the computational horizon for disaster-checking.

Humans seem to do very well at recognizing the need to check for *global* consequences by perceiving *local* features of an action. It remains to be seen whether this characteristic of 1014 x 200 Hz synapses can be duplicated in N 2 Ghz CPUs. See 5.2.2 Layered Mistake Detection.

perfectly Friendly AI. Defined in 2 Challenges of Friendly AI.

A Friendly AI which never makes a mistake and always takes the maximally Friendly action. A purely theoretical concept for infrahuman AIs, though a superintelligence may turn out to be "perfect for all practical purposes."

personal power. Defined in 5.2 Generic Goal Systems.

Power that can be used irrespective of ends, context-insentively. Gandhi had context-sensitive power—many people believed in him, but his power was founded on their respect of his peacefulness. Rockefeller had context-insensitive power; the money that he spent on philanthropy could have been spent on anything else. Humans have an evolved bias which causes them to preferentially accumulate context-insensitive personal effectiveness when they attempt to advance some higher cause. See 5.2.5.1 Anthropomorphic Ethical Injunctions.

peta. 10^{15}. One million billion; a thousand "tera."

philosophical crisis. Defined in 5.6 Interlude: Philosophical Crises.

A "philosophical crisis" is hard to define. I usually think of a "philosophical crisis" as the AI stumbling across some fact that breaks ver loose of the programmers—i.e., the programmers have some deeply buried unconscious prejudice that makes them untrustworthy, or the AI stumbles across a deliberate lie, or the AI discovers objective morality, or something like that—possibly causing all the programmer affirmations to break down simultaneously. The basic notion is that the AI enters totally uncharted territory—something completely orthogonal to everything the programmers ever thought about—or that the AI suddenly realizes that *all* the programmers' actions, right back to vis birth, can't be trusted (this is a *transhuman* AI confronting the realization, one hopes; otherwise this almost certainly indicates a failure of Friendliness). An example would be an AI built solely with external reference semantics confronting the need for shaper/anchor semantics, and so on.

In short, a philosophical crisis is just like a catastrophic failure of Friendliness, except that the AI is *right*—a normative human examining the reasoning would find no errors, and—if sufficiently intelligent to understand what was going on—would start experiencing a similar philosophical crisis.

Most of my readers have probably run through several philosophical crises in the course of growing up. Causal validity semantics, wisdom tournaments, and the underlying Singularity paradigm of increased smartness, are the three major hopes for a Friendly AI being able to handle philosophical crises.

prehuman AI. An AI of below-human ability and intelligence. May refer to an infant-like or tool-level AI; to an AI that only implements a single facet of cognition or that is missing key facets of cognition; or to a fairly mature AI which is still substantially below human level, although "infrahuman" is more often used to describe the latter.

procedural. See declarative.

probabilistic supergoal content. Defined in 5.5.1 External Reference Semantics.

A facet of external reference semantics. Supergoals that a mind can conceive of as being "wrong" or "incomplete." Where supergoals are absolutely certain—"correct by definition"—the AI has a motive to resist any attempt to change the supergoals. See external reference semantics, 5.5.1.1 Probabilistic Supergoal Content, and 5.5.1 External Reference Semantics.

The Psychological Foundations of Culture. The best explanation of evolutionary theory I have ever read is "The Psychological Foundations of Culture" by Tooby and Cosmides (1992), in *The Adapted Mind* by Barkow, Cosmides, and Tooby (1992). (See my personal Bookshelf.) I did not understand evolution until I read this article. Reading this article is a wonderful experience; the article traces the entire chain of causality with absolute, diamond-like precision. It's not perfect, but it's very, very close.

qualia. The substance of conscious experience. "Qualia" is the technical term that describes the redness of red, the mysterious, indescribable, apparently irreducible quality of redness that exists above and beyond a particular frequency of light. If a JPEG viewer stores a set of red pixels, pixels with color 0xFF0000, does it see red the way we do? No. Even if a program simulated all the feature-extraction of the human visual modality, would it actually *see red?*

I first "got" the concept of qualia on reading the sentence "You are not the person who speaks your thoughts; you are the person who hears your thoughts."[101]

See "Facing Up to the Problem of Consciousness" by David Chalmers for a more extensive definition.

qualitative. Defined in Yudkowsky (2001, § 2.1: World-model).

Qualitative properties are selected from a finite set; for example, the binary set of {on, off}, or the eighty-eight member set of "piano keys". A *qualitative match* is when two qualities have identical values. A *qualitative binding* is when the two qualities are hypothesized to be bound together—if, for example, the same item of the 88-member set "piano keys" occurs in two instances. A qualitative binding is the weakest type of binding, since it can often occur by sheer coincidence. However, the larger the set, the less likely a coincidence. Small integers are usually qualitative properties; large integers should be treated as quantitative. See also SPDM, quantitative binding, and structural binding.

quantitative. Defined in Yudkowsky (2001, § 2.1:World-model) .

101. I don't remember where I first heard this, but my guess is Raymond Smullyan.

Quantitative characteristics occupy a continuous range; they are selected from a range of real (i.e., floating-point) numbers. A *quantitative match* is when two quantities are identical. A *quantitative binding* occurs when two or more quantitative variables are equal to sufficient precision that coincidence is effectively impossible. Quantitative bindings can also be established by covariance or other quantitative relations. See also SPDM, qualitative, structural.

reductholism. A word appearing in Gödel, Escher, Bach. "Reductholism" is a synthesis of reductionism and holism; I use it to indicate the general theory of systems with multiple levels, including both the holistic disciplines of looking up and the reductionist disciplines of looking down. See also reductionism and holism.

reductionism. Reductionism: The attitude that the whole is the sum of the parts. To take the reductionist view is to look downward, focus on the low-level elements and the rules governing their interactions. See holism and reductholism. See also Gödel, Escher, Bach: An Eternal Golden Braid, particularly the dialogues "Prelude" and "Ant Fugue".

"No one in his left brain could deny reductionism." —GEB

reflection. Defined in Yudkowsky (2001, § 2.4 Thoughts) .

The ability of a thinking system to think about itself. A reflective mind is one that has an image of itself; a self-model. In Friendly AI, a reflective goal system is one that can regard its own components and content as desirable or undesirable. In mundane programming, a "reflective" programming language is one in which code can access information about code (for example, obtaining a list of all the methods or properties of an object).

relevance horizon. The Universe goes on forever, and what we can say about it goes on even longer. Outside the "relevance horizon" lie all the knowledge and heuristics and skills that are not relevant to the task. See computational horizon.

Humans and AIs probably have *very different* relevance horizons.

Riemann Hypothesis Catastrophe. Defined in 5.2 Generic Goal Systems.

A "failure of Friendliness" scenario in which an AI asked to solve the Riemann Hypothesis turns all the matter in the solar system into computronium, exterminating humanity along the way. (As I recall, I first heard this version—with a slightly different phrasing—from Marvin Minsky.) In variant forms, a subclass of the subgoal stomp error, the Devil's Contract problem (both diabolic and golemic), and the "emergent pressures" scenario *a la* "The Two Faces of Tomorrow" (Hogan 1979)(see 5.2.7 Emergent Phenomena in Generic Goal Systems).

RNUI. Defined in Yudkowsky (2001, § Interlude: Represent, Notice, Understand, Invent) .

Represent, Notice, Understand, Invent. First your AI has to *represent* something, then it has to *notice* it, then it has to *understand* it, then *invent* it. You can't take these items out of sequence.

Representing means having the static data structures to hold the information.

Noticing means being able to see simple relations, to perceive internal coherence; to tell the difference between a representation that makes sense, and a representation composed of random numbers.

Understanding means being able to see goal-oriented properties, and how the thing understood fits into the larger structure of the Universe—the thing's functionality, the causes of that thing's characteristics, and so on.

Inventing means being able to start with a high-level goal—"rapid transportation"—and design a bicycle.

rod logic. A mechanical nanocomputer built using diamondoid rods of a few thousand atoms each. Even though messages can only move at the speed of sound in diamond (~17 km/s == ~$6 \times 1e^{-5}c$), and the calculations in Nanosystems (Drexler 1992) assume ~12 km/s, the very small size of the components would enable an individual rod-logic CPU containing 10^6 transistor-like rod-logic interlocks to operate at 1 GHz clock speeds, executing instructions at ~1000 MIPS. The power consumption for a 1 GHz CPU is estimated to be ~60 nW. The calculations are performed for 300 Kelvin (room temperature). The error rate is $1e^{-64}$ per transistor operation, effectively negligible. (However, the half-life against radiation damage for an unshielded CPU in Earth ambient background radiation is only ~100 years.)

The usual summary is that a one-kilogram, one-cubic-centimeter nanocomputer can contain 10^{12} nanocomputers, consume 100 kW (and dissipate 100 kW of heat—cooling systems are also described), to deliver 10^{21} instructions per second (10^{15} MIPS, a thousand billion billion operations per second). The overall system has a clock speed ~10^6 times faster than the maximum firing rate of a biological neuron, and delivers total computing capacity ~10^4 times the upper-bound estimate for the human brain (~10^{14} synapses operating at ~200 Hz == ~10^{17} ops/second).

There are more speculative *electronic* nanocomputer schemas that would allow ~10^{25} operations per second; also, assuming a parallel-CPU architecture may be conservative when dealing with seed AIs. However, rod logics are easy to analyze and provide a definite lower bound on the computing speeds achievable with molecular manufacturing technology.

rule of derivative validity. Defined in 5.6.2 Causal Validity Semantics.

"Effects cannot have greater validity than their causes." An enormously powerful, fundamentally flawed rule used by the human mind to reason about beliefs and goals.

scalar. A single number, as opposed to two or more numbers. The *speed* of an airplane is a scalar quantity; it can be described by a single number. The *velocity* of an airplane, which includes the direction as well as the speed, is a vector—it must be described by two numbers. (Three numbers, if it's a three-dimensional airplane.)

There's a timeworn joke about mosquitoes and mountain climbers that is usually mentioned at this point, but forget it.

search trees. One of the most venerable tools of AI. In a game of tic-tac-toe, you can make any of nine possible moves, then I can make any of eight possible moves, then you can make any of seven possible moves . . . The computational representation of the game—in a classical AI—would look like a tree; a single node representing the start of the game, with nine branches leading to nine first-move nodes; each first-move node would have eight branches leading to a total of seventy-two possible second-move nodes, and so on. By searching through the entire tree, a classical AI could play a perfect game of tic-tac-toe.

It is possible, even likely, that human cognition involves the use of similar (although much messier) search trees. Or not.

seed AI. Defined in Yudkowsky (2001, § 1.1 Seed AI).

An AI designed for self-understanding, self-improvement, and recursive self-improvement. See Yudkowsky (2001, § 1.1 Seed AI), or the introductory article Yudkowsky (2001, § What is Seed AI?) on the Singularity Institute's website.

selection pressure. A feature of the internal or external environment which causes differential rates of survival and reproduction for different heritable characteristics, including characteristics such as long legs, cognitive processes, and emotions. Any problem with better and worse solutions is technically a selection pressure, but use of the phrase "selection pressure" usually implies that this "problem" is an environmental feature present with sufficient reliability to influence the survival/reproduction rates of multiple individuals across multiple generations. See also complex functional adaptation.

sequitur. An emotional sequitur is an emotion or feeling that automatically binds to certain thoughts or cognitive structures; a logical or intuitive sequitur is a conclusion which is immediately perceived given the trigger of a thought or cognitive

structure. The human brain runs on sequiturs. A sequitur is not quite the unit of a stream of consciousness—the unit of a stream of consciousness is a sentence. Rather, a sequitur covers all the nonverbal conclusions, the perceptual or automatic leaps. "Sequitur" is a term that I find extremely useful, so I hope you aren't annoyed overmuch by my use of it.

sensory modality. Defined in Yudkowsky (2001, § 2.2 Sensory Modalities).

A sensory modality, in an AI, is a module analogous to the human visual cortex, the human auditory cortex, or some other chunk of neurology underlying one of the senses. A modality contains the data structures needed to represent the target domain; the active processing which enables the perception of higher-level features and coherence in that domain; and the interface to the concept level which enables the abstraction of, and visualization of, patterns and objects in that domain.

sexual selection. Sexual selection is a kind of evolutionary positive feedback that can result in ridiculous and even suicidal traits becoming overwhelming evolutionary advantages, leading in some cases to a kind of genetic suicide.

Suppose that there's some species—let's call it a "tailbird"—that happens to have a small, ordinary, unassuming tail. It also happens that the tails of healthy tailbirds are slightly more colorful, more lustrous, then the tails of tailbirds that are sick, or undernourished. One day, a female tailbird is born with a mutation that causes it to sexually prefer tailbirds with bright-colored tails. This is a survival trait—it results in the selection of healthier male mates, with better genes—so the trait propagates until, a few dozen generations later, the entire species population of female tailbirds prefers bright-colored tails.

Now, a male is born that has a *very* bright tail. It's not bright because the male is healthy; it's bright because the male has a mutation that results in a brighter tail. All the females prefer this male, so the mutation is a big success.

This male tailbird isn't actually healthier. In fact, this male is pretty sick. More of his biological resources are going into maintaining that flashy tail. So you might think that the females who preferred that male would tend to have sickly children, and the prefer-bright-tails trait would slowly fade out of the population.

Unfortunately, that's not what happens. What happens is that even though the male has sickly children, they're sickly children with bright tails. And those children also attract a lot of females. Genes can't detect "cheating" and instantly change tactics; that's a monopoly of conscious intelligence. Any females who prefer the non-bright-tailed males will actually do worse. These "wiser" females will have children who are, sexually, out of fashion. Bright tails are no longer a *survival* advantage, but they are a very strong *sexual* advantage.

Selection pressures for sexual advantages are often much stronger than selection pressures for mere survival advantages. From a design perspective this is stupid—but evolution doesn't care. Sexual selection is also a Red Queen's Race (Ridley 1994): It involves competition with conspecifics, so you can never have a tail that's "bright enough." This is how you get peacocks.

Any time you see an enormously exaggerated trait—the tail of a peacock, the antlers of a buck, the intelligence of a human—it's a good guess that two forces are at work: Competition with conspecifics, and sexual selection.

shaper. Defined in 5.6.1 Shaper/Anchor Semantics.

See shaper/anchor semantics. A shaper is one of the philosophical affectors which provide supergoal content. Shapers themselves do not have supergoal status; learning more about shapers is a subgoal of the currently unknown supergoal content (see external reference semantics).

shaper/anchor semantics. Defined in 5.6.1 Shaper/Anchor Semantics.

Humans can achieve philosophical growth because we have a complex, tangled network of mutually interacting affectors. "Shaper/anchor semantics" are a less messy way to duplicate the necessary functionality for philosophical growth in an AI. Shapers affect each other, and can specify supergoal content, but are not themselves supergoals. In terms of the goal architecture, the AI's modeling of shapers is a subgoal of further specifying the supergoal, supported by the "unknown" portion of the supergoal content at any given point—shapers are not are not meta-supergoals.

If the super-supergoal is producing supergoal definitions that maximally satisfy humans, then the system can short-circuit by directly modifying human brain states for maximal satisfaction (for example). If a *shaper* stated that supergoals should maximally satisfy humans—this isn't a real shaper, though it's a good heuristic—then the *current* supergoal content about volitional consent can still veto the idea of nonconsensually modifying the brain state of the programmer, because shapers aren't supergoals and don't have the power to contradict supergoals. Shapers define supergoals. "Shaper-pondering behaviors" are subgoals about how to build the supergoals; they may be behaviors that *affect* supergoals, or rather the current description of the supergoals, but such behaviors still can't *override* supergoals, no more than any other subgoal can. The AI's task of modeling shapers and modifying supergoals accordingly is supported by the currently unknown supergoal content, and can't randomly stomp other supergoal content in the same way that a "meta-supergoal" could stomp supergoals. (Note: This is only a rough explanation.)

An "anchor" is a feature of AI philosophy that reflects the fact that the programmers themselves may not understand their own tangled goal system. An "anchor"

is a probable output of the shaper system whose causal antecedents are not fully known; an "anchor" happens when a human says "I think X is a good idea," but isn't clear about *why*, or isn't sure he/she is fully self-aware of all the valid factors that played a part in the decision. An anchor is an external reference to whichever causes were involved in the decision; it refers to whichever factual cognitive events were occurring within the human's brain. An anchor is a pointer for the AI to learn, by examination of the humans, about a portion of the shaper system that the humans themselves are fuzzy about. See anchor.

Singularity. The transition in human history marked by the technological creation of greater-than-human intelligence; the end of the era of "strictly human intelligence" that has embraced the last fifty thousand years. If one were to draw a single line, dividing into two parts the history of Earth-originating intelligent life, the line would be drawn at the Singularity. See "An Introduction to the Singularity."

Singularity-safed Defined in 5.7 Developmental Friendliness.

A system such that the immediate start of a hard takeoff does not signal automatic catastrophe. Broadly speaking, there are three classes of Singularity-safed AIs:

- A mature AI with structurally complete Friendliness semantics and the threshold level of Friendliness content, such that the acquisition of transhuman intelligence will probably enable the AI to complete the Friendship system with no further inputs.
- A young AI with an ethical injunction telling ver to slow down and ask for directions in case of a hard takeoff.
- Any recursively self-improving AI, no matter how primitive, that has a programmatic "controlled ascent" feature which triggers a save-and-stasis if improvement suddenly accelerates.

See 5.8 Singularity-Safing ("In Case of Singularity, Break Glass").

SPDM. Defined in Yudkowsky (2001, § 2.1 World Model).

Sensory, Predictive, Decisive, Manipulative.

Intelligence is an evolutionary advantage because it enables us to model, predict, and manipulate reality. This idea can be refined into describing four levels of *binding* between a model and reality.

A *sensory* binding is simply a surface correspondence between data structures in the model and whatever high-level properties of reality are being modeled.

A *predictive* binding is one that can be used to correctly predict future sensory inputs.

A *decisive* binding is one that can be used to decide between limited sets of possible actions based on the utility of the predicted results.

A *manipulative* binding is one that can be used to start from a specified result and plan a sequence of actions that will bring about the desired result.

See also qualitative, quantitative, structural.

structural. Defined in Yudkowsky (2001, § 2.1 World-model).

Structural characteristics are made up of multiple qualitative or quantitative components. A *structural match* is when two complex patterns are identical. A *structural binding* occurs when two complex patterns are identical, or bound together to such a degree that coincidence is effectively impossible—only a pattern copy of some kind could have generated the identity. A structural binding is usually the strongest form. See also SPDM, qualitative, quantitative.

subgoal. Defined in 3 An Introduction to Goal Systems.

An event or world-state whose desirability is contingent on its predicted outcome. An action taken in order to achieve some other goal. A thing not desirable in itself, but rather desirable for some other end.

If the "parent goal" is a wet sponge, then placing a dry sponge in water is a "child goal" of getting a wet sponge. The parent goal, in turn, is probably a child goal of something else. "Parent goal" and "child goal" describes the relation between two individual goals, while "supergoal" and "subgoal" describes the general distinction between things intrinsically desirable, and things which are desirable only because they are predicted to lead to other things.

It is very important to distinguish between supergoals and subgoals in Friendly AI, since subgoals are directly and entirely dependent on the current world-state.

subgoal stomp. Defined in 5.2 Generic Goal Systems.

A "failure of Friendliness" scenario in which a subgoal stomps on a supergoal—for example, putting on your shoes before your socks, or turning all the matter in the Universe into computronium because some (ex-)petitioner asked you to solve the Riemann Hypothesis.

subjective. Existing within the mind, as opposed to external reality.

In the strict sense, virtually all things except quarks are subjective. See objective.

In the nonabsolute sense, "subjective" is often used to indicate a statement which is subject to observer bias or speaker deixis, or which makes visible reference to structures that exist only within the mind.

subjunctive. A what-if scenario, e.g. "What if you were to drop that glass of milk?" Something imagined or visualized that *isn't* supposed to describe present or past reality, although a sufficiently attractive what-if scenario might later be turned into a plan, and thence into reality. See also counterfactual.

supergoal. Defined in 3 An Introduction to Goal Systems.

Supergoal content determines which states of the world are seen as *intrinsically* desirable. Opposed to "subgoal", a state desirable only as a means to an end. (The terms"parent goal" and "child goal" are used to refer to local relations between goals. Goal A may be a child goal of B, which in turn is a child goal of C, making B the "parent goal" of A and the "child goal" of C. By contrast, the distinction between supergoal and subgoal is a distinction in kind, at least under the *CFAI* architecture.) See also subgoal.

superintelligence. A level of ability enormously in advance of that possessed by humans. Despite common usage, there is probably no such thing as a "superintelligent AI"; a true superintelligence would no more resemble an AI than a human resembles an amoeba. Singularity analysis commonly distinguishes between *weak superintelligence,* which is human thought at enormously higher speeds, and *strong superintelligence,* involving an intelligence difference similar to that distinguishing a human from a chimpanzee, or a dog, or a possibly a rock.

It has been suggested[102] that "transhuman" should refer to a mind with perhaps three or four times the computing capacity (or cognitive capacity) of a human, that "superintelligence" should refer to a mind with perhaps a hundred or a thousand times human capacity, and that "Power" (a la Vinge) should refer to a mind with billions or trillions of times the computing power of Earth's entire current population. It is noteworthy that even a transhuman is still smarter than we are and thus fundamentally incomprehensible, and also that a desktop rod logic suffices for superintelligence under this definition.

supersaturated Friendliness. Defined in 5.7 Developmental Friendliness.

A Friendly AI that has as much Friendliness structure and Friendliness content as ve can usefully represent (for use then or later) is supersaturatedly Friendly.

102. I cannot recall who originally made the suggestion.

Supersaturated Friendliness is achieved by adding ethical injunctions long before they're needed, letting the AI avoid anything that looks the least bit unFriendly "just to be sure," trying to implement the last 10% of functionality that takes 90% of the effort, and doing it all two or more steps in advance of the development stage where it becomes unavoidably necessary.

Sysop. Defined in 2 Challenges of Friendly AI.

A superintelligent AI which provides an "operating system" for all the matter in the Solar System. A living peace treaty with the power to enforce itself. A friendly superintelligence ready to do whatever the "inhabitants" ask—including upgrading us to superintelligence ourselves—so long as it doesn't violate the citizenship rules derived from Friendliness. (For example, asking the Sysop to kill someone else, or to use more than your share of the Solar System's resources, or even to upload and upgrade someone against their will.) One of the possible outcomes of the Singularity; see 5.9 Interlude: Of Transition Guides and Sysops.

tera. 10^{12}. One thousand billion; a thousand "giga."

transhuman. A level of ability substantially in advance of that possessed by humans. A transhuman AI is one which has substantially exceeded human intelligence. A further distinction can be made between *weak transhumanity,* which involves intelligence differentials roughly of the same order as those that distinguish Einstein from the average human, and *strong transhumanity,* involving an intelligence difference similar to that distinguishing a human from a chimpanzee. See also the definitions of weak and strong superintelligence.

Transition Guide. Defined in 2 Challenges of Friendly AI.

For the last fifty thousand years, humanity has been essentially at the mercy of a hostile universe. Post-Singularity, things will be different. A Transition Guide is a Friendly seed AI that has achieved superintelligence and created nanotechnology, or whatever other tools are needed to (a) perceive whatever state of existence comes after the Singularity and (b) handle the transition. A Transition Guide is the bridge between modern-day humanity and the other side of a Friendly Singularity.

truly perfect Friendly AI. Defined in 2 Challenges of Friendly AI.

An AI which always takes the perfectly Friendly action, *even if you programmed it to do something else.* A truly perfect Friendly AI has sufficient "strength of philosophical personality"—while still matching the intuitive aspects of friendliness, such as not killing off humans and so on—that we are more inclined to trust the philosophy of the Friendly AI, than the philosophy of the original programmers.

The real-world analogue would be a "philosophically human-equivalent" or "philosophically transhuman" Friendly AI.

turing-computability. If you really haven't heard the term "Turing-computable" before, the first thing you need to do is read Douglas R. Hofstadter's Gödel, Escher, Bach: An Eternal Golden Braid. Drop whatever you're doing, get the book, and read it. It's no substitute, but there's also a nice definition of "Turing machine" in the Stanford Encyclopedia of Philosophy; also, I give a sample visualization of a Turing machine in *Unimplemented section: Causality*.

Any modern digital computer can, in theory, be simulated by a Turing machine. Any modern computer can also simulate a Turing machine, at least until it runs out of memory (Turing machines, as mathematical concepts, have an infinite amount of memory). In essence, Turing demonstrated that a very wide class of computers, including modern Pentiums and PowerPC chips, are all fundamentally equivalent—they can all simulate each other, given enough time and memory.

There is a task known as the *halting problem*—in essence, to determine whether Turing machine X, acting on input Y, will halt or continue forever. Since the actions of a Turing machine are clearly defined and unambiguous, the halting problem obviously has a true, unique, mathematically correct yes-or-no answer for any specific question. Turing, using a diagonalization argument, demonstrated that no Turing machine can solve the general halting problem. Since any modern digital computer can be simulated by a Turing machine, it follows that no digital computer can solve the halting problem. The halting problem is *noncomputable.* (This doesn't necessarily demonstrate a fundamental "inferiority" of computers, since there's no reason to suppose that *humans* can solve the halting problem. In fact, we can't do so simply by virtue of the fact that we have limited memories.)

A controversial question is whether *our Universe* is Turing-computable—that is, can the laws of physics be simulated to arbitrary accuracy by a digital computer with sufficient speed and memory? And if not, do the uncomputabilities carry over to (a) qualia and (b) intelligence? I don't wish to speculate here about qualia; I don't understand them and neither do you. However, I do understand intelligence, so I'm fairly sure that even if qualia are noncomputable, that noncomputability doesn't carry over into human general intelligence.

I happen to believe that our physical Universe *is* noncomputable, mostly because I don't trust the Turing formalism. The concept of causality involved strikes me as fundamentally subjective; there's no tail-end recursion explaining why anything exists in the first place; I've never seen a good mathematical definition of "instantiation"; plus a lot of other reasons that are *waay* beyond the scope of this

document. If the physical Universe is noncomputable, I believe that qualia are probably noncomputable as well. However, this belief is strictly in my personal capacity and is not defended (or attacked, or discussed) in GISAI. I am in the extreme minority in so believing, and in an even smaller minority (possibly a minority of one) in believing that *qualia* are noncomputable but that this says nothing about the impossibility, or even the difficulty, of achieving *real intelligence* on computable hardware. I don't believe that semantics require special causal powers, that Gödel's Theorem is at all difficult to explain to a computable intelligence, that human mathematical ability is noncomputable, that humans are superior to mere computers, or any of the other memes that customarily go along with the "noncomputability" meme.

unFriendly. An unFriendly action: an action which violates Friendliness. An unFriendly AI: an AI which is not designed using Friendly AI design principles, or an AI which has undergone a catastrophic failure of Friendliness. Intuitively: The AI starts shooting people (if near-human) or exterminates humanity and dissassembles us for spare atoms (if transhuman).

unity of will. Defined in 5.3 Seed AI Goal Systems.

See moral deixis and 5.3.3 Unity of Will. The ability of Friendship programmers and AIs to cooperate, even against the AI's "interests" (from a human perspective, at least), because their wills and decisions are sufficiently close in the near-term, and sufficiently convergent in the long-term.

Given perfect identity of knowledge, supergoals, and cognitive processes, the result is perfect unity of will. Small variances in knowledge resulting from different sensory viewpoints should not be sufficient to break the characteristic functionality of "unity of will," unless the variance results in a desirability differential greater than the differential desirability of the "unity of will" process. See 5.3.3 Unity of Will for more details.

upload. Colloquially, a human being that has been uploaded into a computer is an "upload." The classic method is a "Moravec Transfer"; a medical nanobot enters the brain, swims up to a neuron, scans the neuron until the outputs can be perfectly predicted (or predicted to within an order of magnitude greater accuracy than the neuron's sensitivity to random external thermal fluctuations), replaces the neuron (slowly or quickly), and takes over that neuron's functional role within the brain—perhaps first enclosing the axons, dendrites, and cell body in a robotic casing, then switching from the cell body's outputs to the nanobot's computed outputs. If this procedure is repeated for each neuron in the brain, a mind can actually be transferred from biological to computational substrate without ever losing conscious-

ness. Presumably hormones, glands, neurotransmitter diffusion, electrical field effects, long-term potentiation, and so on, would need to be simulated as well. But a neuron is a *lot* larger than a medical nanobot, so there's plenty of room to scan right down to the individual microtubule dimers.

If human neurons are noncomputable *a la* Hameroff and Penrose, then the nanobots and the new computational substrate need to employ noncomputable hardware as well. In essence, the mind would be transferred to programmable high-speed "superneurons" rather than silicon transistors as such.

ve. A gender-neutral pronoun. The equivalent of "he" or "she."

ver. A gender-neutral pronoun. The equivalent of "him" or "her."

verself. A gender-neutral pronoun. The equivalent of "himself" or "herself."

vis. A gender-neutral pronoun. The equivalent of "his" or "her" (or "hers").

wirehead. Defined in 4 Beyond Anthropomorphism.

Larry Niven (1995) once wrote a story, "Death by Ecstasy" , dealing with the possibility of humans wiring their pleasure centers with electrodes—a habit he called "wireheading." Another story that deals with the issue of wireheading is the novel *Mindkiller* by Spider Robinson (1982). Hence the term.

In Friendly AI terms, a "wirehead AI" is one that has substituted the indicator of success or desirability for the supergoal content itself. The "wirehead fallacy" is the idea that any sufficiently intelligent entity necessarily values pleasure above all else. The "wirehead problem" is any failure of Friendliness in which an internal indicator becomes the de facto supergoal; "wirehead problem" also refers to any failure of the goal architecture in which some piece of mindstuff "cheats" and assigns itself desirability (this problem turned up *fifteen years ago* in Doug Lenat's Eurisko; see 5.3.5 FoF: Wireheading 2).

wisdom tournament. Defined in 5.3 Seed AI Goal Systems.

One of the major sources of human wisdom is our need to compensate for our built-in errors. We invented science to deal with politically biased belief systems. We invented altruism to deal with our selfishness. We invented ethics to deal with our social observer-biased rationalizations. A wisdom tournament is a subjunctive, self-simulated version of the AI that shows the AI what the AI "would have done" if the AI had been dumber, slower, biased. Without introducing actual errors, the AI gains reliable information about what would have happened if the AI had been running on unreliable simulated hardware, with random perturbations

to the software, ethical misinformation, factual misinformation, tempting ends-justify-the-means scenarios, and an instinct to kill and destroy. If the Friendliness, and the heuristics the AI has learned from past tournaments, are enough to walk through all that without making a single major error, then the Friendliness may be strong enough to face real life.

See 5.3.4 Wisdom Tournaments.

7.4. Version History

June 15, 2001. CFAI 1.0. Released along with *SIAI Guidelines on Friendly AI.* 909K.

June 12, 2001. CFAI 0.9.05. Broke off 5.6 Interlude: Philosophical Crises as separate section. Content added to 5.8 Singularity-Safing ("In Case of Singularity, Break Glass"). Content added to 5.5.1 External Reference Semantics. Content added to 5.1.1 Cleanly Causal Goal Systems. 909K.

June 10, 2001. CFAI 0.9.04. More bugfixes. Old "Friendly goal systems" renamed to "Friendship structure"; associated folder renamed from "CFAI/design/friendly/" to "CFAI/design/structure/." 5.5 Interlude: Why Structure Matters moved to inside 5.4 Friendship Structure. 5.1.1 Cleanly Causal Goal Systems added; much content moved there from 5.5.1 External Reference Semantics, 5.3 Seed AI Goal Systems and elsewhere. 860K. Not uploaded.

June 7, 2001. CFAI 0.9.03. "Design requirements of Friendly AI" renamed to "Challenges of Friendly AI." "An Introduction to Goal Systems" split off into separate page. Numerous bugfixes. 822K. Not uploaded.

May 18, 2001. CFAI 0.9.02. Split the original document, "Coding a Transhuman AI," into *General Intelligence and Seed AI* and *Creating Friendly AI*. Minor assorted bugfixes. 782K.

Apr 24, 2001. CFAI 0.9.01. Uploaded printable version. Some minor suggested bugfixes.

Apr 18, 2001. CFAI 0.9. "Open commentary" version of *Creating Friendly AI* begins circulating in transhumanist community. Linked in to primary Singularity Institute website. 748K. Changed copyright to "2001" and "Singularity Institute" instead of legacy "2000" and "Eliezer Yudkowsky."

Apr 09, 2001. CFAI 0.9a. Vast amounts of new material. 5.4 Friendship Structure split into multiple subpages. Still not linked in. 744K.

Mar 21, 2001. CFAI 0.4. Finished 5.3 Seed AI Goal Systems; still not linked in. 437K.

Feb 14, 2001. CFAI 0.3. Large amounts of new material added. Still not linked in. 363K.

Jan 16, 2001. CFAI 0.2. Placed pre-publication multi-page version of "Coding a Transhuman AI" online to include the in-progress section "Friendly AI." 187K. Not published or linked-in to Singularity Institute website.

References

Asimov, Isaac. 1947. "Little Lost Robot." *Astounding Science-Fiction,* March, 111–132.

Barkow, Jerome H., Leda Cosmides, and John Tooby, eds. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. `http://www.jetpress.org/volume9/risks.html`.

Drexler, K. Eric. 1986. *Engines of Creation.* Garden City, NY: Anchor.

———. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* New York: Wiley.

Egan, Greg. 1998. *Diaspora.* 1st ed. New York: HarperPrism.

Freitas Jr., Robert A. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." Foresight Institute. April. Accessed November 24, 2012. `http://www.foresight.org/nano/Ecophagy.html`.

Hogan, James P. 1979. *The Two Faces of Tomorrow.* 1st ed. New York: Ballantine Books.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. `http://www.wired.com/wired/archive/8.04/joy.html`.

Konopinski, Emil J., C. Marvin, and Edward Teller. 1946. *LA-602: Ignition of the Atmosphere with Nuclear Bombs.* Technical report. Los Alamos National Laboratory, Los Alamos, NM. `http://library.lanl.gov/la-pubs/00329010.pdf`.

Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1-2): 61–98. doi:`10.1016/S0004-3702(83)80005-8`.

Machiavelli, Niccolò. (1532) 1998. *The Prince.* 2nd ed. Edited by Harvey C. Mansfield. University of Chicago Press.

Niven, Larry. 1995. "Death by Ecstasy." In *Flatlander,* 1–70. New York: Ballantine Books.

Pratchett, Terry. 1996. *Feet of Clay: A Novel of Discworld.* Discworld Series. New York: HarperTorch.

Ridley, Matt. 1994. *The Red Queen: Sex and the Evolution of Human Nature.* 1st ed. New York: Macmillan.

Robinson, Spider. 1982. *Mindkiller: A Novel of the Near Future.* 1st ed. New York: Holt, Rinehart & Winston.

Tooby, John, and Leda Cosmides. 1992. "The Psychological Foundations of Culture." In Barkow, Cosmides, and Tooby 1992, 19–136.

Tversky, Amos, and Daniel Kahneman. 1986. "Rational Choice and the Framing of Decisions." In "The Behavioral Foundations of Economic Theory." Supplement, *Journal of Business* 59 (4, pt. 2): S251–S278. `http://www.jstor.org/stable/2352759`.

Wilson, Robert A., and Frank Keil, eds. 1999. *The MIT Encyclopedia of the Cognitive Sciences.* 1st ed. Bradford Books. Cambridge, MA: MIT Press.

Yudkowsky, Eliezer. 2001. *General Intelligence and Seed AI.* Version 2.3. The Singularity Institute, San Francisco, CA. `http://intelligence.org/files/GISAI.html`.