Artificial Intelligence as a Positive and Negative Factor in Global Risk

Dr Prof Engr Mr Santosh Kumar
Senior Technical Officer, Hindustan Aeronautics Limited

Indian Institute of Science (Research University),Bengaluru, Karnataka,India

1. Introduction

By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it. Of course this problem is not limited to the field of AI. Jacques Monod wrote: "A curious aspect of the theory of evolution is that everybody thinks he understands it" (Monod 1975). My father, a physicist, complained about people making up their own theories of physics; he wanted to know why people did not make up their own theories of chemistry. (Answer: They do.) Nonetheless the problem seems to be unusually acute in Artificial Intelligence. The field of AI has a reputation for making huge promises and then failing to deliver on them. Most observers conclude that AI is hard; as indeed it is. But the *embarrassment* does not stem from the difficulty. It is difficult to build a star from hydrogen, but the field of stellar astronomy does not have a terrible reputation for promising to build stars and then failing. The critical inference is *not* that AI is hard, but that, for some reason, it is very easy for people to think they know far more about Artificial Intelligence than they actually do.

In my other chapter for *Global Catastrophic Risks*, "Cognitive Biases Potentially Affecting Judgment of Global Risks" (Yudkowsky 2008), I opened by remarking that few people would deliberately choose to destroy the world; a scenario in which the Earth is destroyed *by mistake* is therefore very worrisome. Few people would push a button that they clearly knew would cause a global catastrophe. But if people are liable to confidently believe that the button does something quite different from its actual consequence, that is cause indeed for alarm.

It is far more difficult to write about global risks of Artificial Intelligence than about cognitive biases. Cognitive biases are settled science; one need simply quote the literature. Artificial Intelligence is not *settled* science; it belongs to the frontier, not to the textbook. And, for reasons discussed in a later section, on the topic of *global catastrophic risks* of Artificial Intelligence, there is virtually no discussion in the existing technical literature. I have perforce analyzed the matter from my own perspective; given my own conclusions and done my best to support them in limited space. It is not that I have neglected to cite the existing major works on this topic, but that, to the best of my ability to discern, there are no existing major works to cite (as of January 2006).

It may be tempting to ignore Artificial Intelligence because, of all the global risks discussed in this book, AI is hardest to discuss. We cannot consult actuarial statistics to assign small annual probabilities of catastrophe, as with asteroid strikes. We cannot use calculations from a precise, precisely confirmed model to rule out events or place infinitesimal upper bounds on their probability, as with proposed physics disasters. But this makes AI catastrophes more worrisome, not less.

The effect of many cognitive biases has been found to *increase* with time pressure, cognitive busyness, or sparse information. Which is to say that *the more difficult the analytic challenge*, the more important it is to avoid or reduce bias. Therefore I *strongly recommend* reading "Cognitive Biases Potentially Affecting Judgment of Global Risks" before continuing with this paper.

2. Anthropomorphic Bias

When something is universal enough in our everyday lives, we take it for granted to the point of forgetting it exists.

Imagine a complex biological adaptation with ten necessary parts. If each of ten genes are independently at 50% frequency in the gene pool—each gene possessed by only half the organisms in that species—then, on average, only 1 in 1024 organisms will possess the full, functioning adaptation. A fur coat is not a significant evolutionary advantage unless the environment reliably challenges organisms with cold. Similarly, if gene B depends on gene A, then gene B has no significant advantage unless gene A forms a reliable part of the *genetic* environment. *Complex, interdependent* machinery is necessarily *universal* within a sexually reproducing species; it cannot evolve otherwise (Tooby and Cosmides 1992). One robin may have smoother feathers than another, but they will both have wings. Natural selection, while feeding on variation, uses it up (Sober 1984).

In every known culture, humans experience joy, sadness, disgust, anger, fear, and surprise (Brown 1991), and indicate these emotions using the same facial expressions (Ekman and Keltner 1997). We all run the same engine under our hoods, though we may be painted different colors; a principle which evolutionary psychologists call the *psychic unity of humankind* (Tooby and Cosmides 1992). This observation is both explained and required by the mechanics of evolutionary biology.

An anthropologist will not excitedly report of a newly discovered tribe: "They eat food! They breathe air! They use tools! They tell each other stories!" We humans forget how alike we are, living in a world that only reminds us of our differences.

Humans evolved to model other humans—to compete against and cooperate with our own conspecifics. It was a reliable property of the ancestral environment that every powerful intelligence you met would be a fellow human. We evolved to understand our fellow humans *empathically*, by placing ourselves in their shoes; for that which needed to be modeled was similar to the modeler. Not surprisingly, human beings often "anthropomorphize"—expect humanlike properties of that which is not human. In *The Matrix* (Wachowski and Wachowski 1999), the supposed "artificial intelligence" Agent Smith initially appears utterly cool and collected, his face passive and unemotional. But later,

while interrogating the human Morpheus, Agent Smith gives vent to his disgust with humanity—and his face shows the human-universal facial expression for disgust.

Querying your own human brain works fine, as an adaptive instinct, if you need to predict other humans. If you deal with any other kind of optimization process—if, for example, you are the eighteenth-century theologian William Paley, looking at the complex order of life and wondering how it came to be—then anthropomorphism is flypaper for unwary scientists, a trap so sticky that it takes a Darwin to escape.

Experiments on anthropomorphism show that subjects anthropomorphize unconsciously, often flying in the face of their deliberate beliefs. In a study by Barrett and Keil (1996), subjects strongly professed belief in non-anthropomorphic properties of God: that God could be in more than one place at a time, or pay attention to multiple events simultaneously. Barrett and Keil presented the same subjects with stories in which, for example, God saves people from drowning. The subjects answered questions about the stories, or retold the stories in their own words, in such ways as to suggest that God was in only one place at a time and performed tasks sequentially rather than in parallel. Serendipitously for our purposes, Barrett and Keil also tested an additional group using otherwise identical stories about a superintelligent computer named "Uncomp." For example, to simulate the property of omnipresence, subjects were told that Uncomp's sensors and effectors"cover every square centimeter of the earth and so no information escapes processing." Subjects in this condition also exhibited strong anthropomorphism, though significantly less than the God group. From our perspective, the key result is that even when people consciously believe an AI is unlike a human, they still visualize scenarios as if the AI were anthropomorphic (but not quite as anthropomorphic as God).

Anthropomorphic bias can be classed as insidious: it takes place with no deliberate intent, without conscious realization, and in the face of apparent knowledge.

Back in the era of pulp science fiction, magazine covers occasionally depicted a sentient monstrous alien—colloquially known as a bug-eyed monster or BEM—carrying off an attractive human female in a torn dress. It would seem the artist believed that a non-humanoid alien, with a wholly different evolutionary history, would sexually desire human females. People don't make mistakes like that by explicitly reasoning: "All minds are likely to be wired pretty much the same way, so presumably a BEM will find human females sexually attractive." Probably the artist did not *ask* whether a giant bug *perceives* human females as attractive. Rather, a human female in a torn dress *is sexy*—inherently so, as an intrinsic property. They who made this mistake did not think

about the insectoid's mind; they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the BEM doesn't enter into it.[1]

People need not realize they are anthropomorphizing (or even realize they are engaging in a questionable act of predicting other minds) in order for anthropomorphism to supervene on cognition. When we try to reason about other minds, each step in the reasoning process may be contaminated by assumptions so ordinary in human experience that we take no more notice of them than air or gravity. You object to the magazine illustrator: "Isn't it more likely that a giant male bug would sexually desire giant female bugs?" The illustrator thinks for a moment and then says to you: "Well, even if an insectoid alien starts out liking hard exoskeletons, after the insectoid encounters human females it will soon realize that human females have much nicer, softer skins. If the aliens have sufficiently advanced technology, they'll genetically engineer themselves to like soft skins instead of hard exoskeletons."

This is a fallacy-at-one-remove. After the alien's anthropomorphic thinking is pointed out, the magazine illustrator takes a step back and tries to justify the alien's conclusion as a neutral product of the alien's reasoning process. Perhaps advanced aliens *could* reengineer themselves (genetically or otherwise) to like soft skins, but would they *want* to? An insectoid alien who likes hard skeletons will not wish to change itself to like soft skins instead—not unless natural selection has somehow produced in it a distinctly human sense of meta-sexiness. When using long, complex chains of reasoning to argue in favor of an anthropomorphic conclusion, each and every step of the reasoning is another opportunity to sneak in the error.

And it is also a serious error to begin from the conclusion and search for a neutral-seeming line of reasoning leading there; this is rationalization. If it is self-brain-query which produced that first fleeting mental image of an insectoid chasing a human female, then anthropomorphism is the underlying cause of that belief, and no amount of rationalization will change that.

Anyone seeking to reduce anthropomorphic bias in themselves would be well-advised to study evolutionary biology for practice, preferably evolutionary biology with math. Early biologists often anthropomorphized natural selection—they believed that evolution would do the same thing they would do; they tried to predict the effects of evolution by putting themselves "in evolution's shoes." The result was a great deal of nonsense,

1. This is a case of a deep, confusing, and extraordinarily common mistake which E. T. Jaynes named the "mind projection fallacy" (Jaynes 2003). Jaynes, a theorist of Bayesian probability, coined "mind projection fallacy" to refer to the error of confusing states of knowledge with properties of objects. For example, the phrase "mysterious phenomenon" implies that mysteriousness is a property of the phenomenon itself. If I am ignorant about a phenomenon, then this is a fact about my state of mind, not a fact about the phenomenon.

which first began to be *systematically* exterminated from biology in the late 1960s, e.g. by Williams (1966). Evolutionary biology offers both mathematics and case studies to help hammer out anthropomorphic bias.

2.1. The Width of Mind Design Space

Evolution strongly conserves some structures. Once other genes evolve which depend on a previously existing gene, that early gene is set in concrete; it cannot mutate without breaking multiple adaptations. Homeotic genes—genes controlling the development of the body plan in embryos—tell many other genes when to activate. Mutating a homeotic gene can result in a fruit fly embryo that develops normally except for not having a head. As a result, homeotic genes are so strongly conserved that many of them are the same in humans and fruit flies—they have not changed since the last common ancestor of humans and bugs. The molecular machinery of ATP synthase is essentially the same in animal mitochondria, plant chloroplasts, and bacteria; ATP synthase has not changed significantly since the rise of eukaryotic life two billion years ago.

Any two AI designs might be less similar to one another than you are to a petunia.

The term "Artificial Intelligence" refers to a vastly greater *space of possibilities* than does the term "Homo sapiens." When we talk about "AIs" we are really talking about *minds-in-general*, or optimization processes in general. Imagine a map of mind design space. In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the *space of minds-in-general*. The entire map floats in a still vaster space, *the space of optimization processes*. Natural selection creates complex functional machinery without mindfulness; evolution lies inside the space of optimization processes but outside the circle of minds.

It is this *enormous* space of possibilities which outlaws anthropomorphism as legitimate reasoning.

3. Prediction and Design

We cannot query our own brains for answers about nonhuman optimization processes—whether bug-eyed monsters, natural selection, or Artificial Intelligences. How then may we proceed? How can we predict what Artificial Intelligences will do? I have deliberately asked this question in a form that makes it intractable. By the halting problem, it is impossible to predict whether an *arbitrary* computational system implements any input-output function, including, say, simple multiplication (Rice 1953). So how is it possible that human engineers can build computer chips which reliably implement multiplication? Because human engineers deliberately use designs that they *can* understand.

Anthropomorphism leads people to believe that they can make predictions, given no more information than that something is an "intelligence"—anthropomorphism will go on generating predictions regardless, your brain automatically putting itself in the shoes of the "intelligence." This may have been one contributing factor to the embarrassing history of AI, which stems not from the difficulty of AI as such, but from the mysterious ease of acquiring erroneous beliefs about what a given AI design accomplishes.

To make the statement that a bridge will support vehicles up to 30 tons, civil engineers have two weapons: choice of initial conditions, and safety margin. They need not predict whether an *arbitrary* structure will support 30-ton vehicles, only design a single bridge of which they can make this statement. And though it reflects well on an engineer who can correctly calculate the exact weight a bridge will support, it is also acceptable to calculate that a bridge supports vehicles of *at least* 30 tons—albeit to assert this vague statement *rigorously* may require much of the same theoretical understanding that would go into an exact calculation.

Civil engineers hold themselves to high standards in predicting that bridges will support vehicles. Ancient alchemists held themselves to much lower standards in predicting that a sequence of chemical reagents would transform lead into gold. How much lead into how much gold? What is the exact causal mechanism? It's clear enough why the alchemical researcher *wants* gold rather than lead, but why should this sequence of reagents transform lead to gold, instead of gold to lead or lead to water?

Some early AI researchers believed that an artificial neural network of layered thresholding units, trained via backpropagation, would be "intelligent." The wishful thinking involved was probably more analogous to alchemy than civil engineering. Magic is on Donald Brown's list of human universals (Brown 1991); science is not. We don't *instinctively* see that alchemy won't work. We don't *instinctively* distinguish between rigorous understanding and good storytelling. We don't *instinctively* notice an expectation of positive results which rests on air.

The human species came into existence through natural selection, which operates through the nonchance retention of chance mutations. One path leading to global catastrophe—to someone pressing the button with a mistaken idea of what the button does—is that Artificial Intelligence comes about through a similar accretion of working algorithms, with the researchers having no deep understanding of how the combined system works.

Nonetheless they believe the AI will be friendly, with no strong visualization of the exact processes involved in producing friendly behavior, or any detailed understanding of what they mean by friendliness. Much as early AI researchers had strong mistaken vague expectations for their programs' intelligence, we imagine that these AI researchers

succeed in constructing an intelligent program, but have strong mistaken vague expectations for their program's friendliness.

Not knowing how to build a friendly AI is not deadly, of itself, in any specific instance, if you know you don't know. It's *mistaken* belief that an AI will be friendly which implies an obvious path to global catastrophe.

4. Underestimating the Power of Intelligence

We tend to see individual differences instead of human universals. Thus when someone says the word "intelligence," we think of Einstein, instead of humans.

Individual differences of human intelligence have a standard label, *Spearman's g* aka *g-factor*, a controversial interpretation of the solid experimental result that different intelligence tests are highly correlated with each other and with real-world outcomes such as lifetime income (Jensen 1998). Spearman's *g* is a statistical abstraction from individual differences of intelligence between humans, who as a *species* are far more intelligent than lizards. Spearman's *g* is abstracted from millimeter height differences among a species of giants.

We should not confuse Spearman's g with *human general intelligence*, our capacity to handle a wide range of cognitive tasks incomprehensible to other species. General intelligence is a between-species difference, a complex adaptation, and a human universal found in all known cultures. There may as yet be no academic consensus on intelligence, but there is no doubt about the existence, or the power, of the thing-to-be-explained. There is *something* about humans that let us set our footprints on the Moon.

But the word "intelligence" commonly evokes pictures of the starving professor with an IQ of 160 and the billionaire CEO with an IQ of merely 120. Indeed there are differences of individual ability apart from "book smarts" which contribute to relative success in the human world: enthusiasm, social skills, education, musical talent, rationality. Note that each factor I listed is *cognitive*. Social skills reside in the brain, not the liver. And jokes aside, you will not find many CEOs, nor yet professors of academia, who are chimpanzees. You will not find many acclaimed rationalists, nor artists, nor poets, nor leaders, nor engineers, nor skilled networkers, nor martial artists, nor musical composers who are mice. Intelligence is the foundation of human power, the strength that fuels our other arts.

The danger of confusing general intelligence with *g*-factor is that it leads to tremendously underestimating the potential impact of Artificial Intelligence. (This applies to underestimating potential good impacts, as well as potential bad impacts.) Even the phrase "transhuman AI" or "artificial superintelligence" may still evoke images of book-smarts-in-a-box: an AI that's *really good* at cognitive tasks stereotypically associated

with "intelligence," like chess or abstract mathematics. But not superhumanly persuasive; or far better than humans at predicting and manipulating human social situations; or inhumanly clever in formulating long-term strategies. So instead of Einstein, should we think of, say, the 19th-century political and diplomatic genius Otto von Bismarck? But that's only the mirror version of the error. The entire range from village idiot to Einstein, or from village idiot to Bismarck, fits into a small dot on the range from amoeba to human.

If the word "intelligence" evokes Einstein instead of humans, then it may sound sensible to say that intelligence is no match for a gun, as if guns had grown on trees. It may sound sensible to say that intelligence is no match for money, as if mice used money. Human beings didn't *start out* with major assets in claws, teeth, armor, or any of the other advantages that were the daily currency of other species. If you had looked at humans from the perspective of the rest of the ecosphere, there was no hint that the squishy things would eventually clothe themselves in armored tanks. We *invented* the battleground on which we defeated lions and wolves. We did not match them claw for claw, tooth for tooth; we had our own ideas about what mattered. Such is the power of creativity.

Vinge (1993) aptly observed that a future containing smarter-than-human minds is *different in kind*. Artificial Intelligence is not an amazing shiny expensive gadget to advertise in the latest tech magazines. Artificial Intelligence does not belong in the same graph that shows progress in medicine, manufacturing, and energy. Artificial Intelligence is not something you can casually mix into a *lumpenfuturistic* scenario of skyscrapers and flying cars and nanotechnological red blood cells that let you hold your breath for eight hours. Sufficiently tall skyscrapers don't potentially start doing their own engineering. Humanity did not rise to prominence on Earth by holding its breath longer than other species.

The catastrophic scenario which stems from underestimating the power of intelligence is that someone builds a button, and doesn't care enough what the button does, because they don't think the button is powerful enough to hurt them. Or, since underestimating the power of intelligence implies a proportional underestimate of the potential impact of Artificial Intelligence, the (presently tiny) group of concerned researchers and grantmakers and individual philanthropists who handle existential risks on behalf of the human species, will not pay enough attention to Artificial Intelligence. Or the wider field of AI will not pay enough attention to risks of strong AI, and therefore good tools and firm foundations for friendliness will not be available when it becomes possible to build strong intelligences.

And one should not fail to mention—for it also impacts upon existential risk—that Artificial Intelligence could be the powerful solution to other existential risks, and by

mistake we will ignore our best hope of survival. The point about underestimating the potential impact of Artificial Intelligence is symmetrical around potential good impacts and potential bad impacts. That is why the title of this chapter is "Artificial Intelligence as a Positive and Negative Factor in Global Risk," not "Global Risks of Artificial Intelligence." The prospect of AI interacts with global risk in more complex ways than that; if AI were a pure liability, matters would be simple.

5. Capability and Motive

There is a fallacy oft-committed in discussion of Artificial Intelligence, especially AI of superhuman capability. Someone says: "When technology advances far enough we'll be able to build minds far surpassing human intelligence. Now, it's obvious that how large a cheesecake you can make depends on your intelligence. A superintelligence could build *enormous* cheesecakes—cheesecakes the size of cities—by golly, the future will be full of giant cheesecakes!" The question is whether the superintelligence *wants* to build giant cheesecakes. The vision leaps directly from *capability* to *actuality*, without considering the necessary intermediate of *motive*.

The following chains of reasoning, considered in isolation without supporting argument, all exhibit the Fallacy of the Giant Cheesecake:

- A sufficiently powerful Artificial Intelligence could overwhelm any human resistance and wipe out humanity. (And the AI would decide to do so.) Therefore we should not build AI.
- A sufficiently powerful AI could develop new medical technologies capable of saving millions of human lives. (And the AI would decide to do so.) Therefore we should build AI.
- Once computers become cheap enough, the vast majority of jobs will be performable by Artificial Intelligence more easily than by humans. A sufficiently powerful AI would even be better than us at math, engineering, music, art, and all the other jobs we consider meaningful. (And the AI will decide to perform those jobs.) Thus after the invention of AI, humans will have nothing to do, and we'll starve or watch television.

5.1. Optimization Processes

The above deconstruction of the Fallacy of the Giant Cheesecake invokes an intrinsic anthropomorphism—the idea that motives are separable; the implicit assumption that by talking about "capability" and "motive" as separate entities, we are carving reality at its joints. This is a useful slice but an anthropomorphic one.

To view the problem in more general terms, I introduce the concept of an *optimization process*: a system which hits small targets in large search spaces to produce coherent real-world effects.

An optimization process steers the future into particular regions of the possible. I am visiting a distant city, and a local friend volunteers to drive me to the airport. I do not know the neighborhood. When my friend comes to a street intersection, I am at a loss to predict my friend's turns, either individually or in sequence. Yet I can predict the *result* of my friend's unpredictable actions: we will arrive at the airport. Even if my friend's house were located elsewhere in the city, so that my friend made a wholly different sequence of turns, I would just as confidently predict our destination. Is this not a strange situation to be in, scientifically speaking? I can predict the *outcome* of a process, without being able to predict any of the *intermediate steps* in the process. I will speak of the region into which an optimization process steers the future as that optimizer's *target*.

Consider a car, say a Toyota Corolla. Of all possible configurations for the atoms making up the Corolla, only an infinitesimal fraction qualify as a useful working car. If you assembled molecules at random, many *many* ages of the universe would pass before you hit on a car. A tiny fraction of the design space does describe vehicles that we would recognize as faster, more efficient, and safer than the Corolla. Thus the Corolla is not *optimal* under the designer's goals. The Corolla is, however, *optimized*, because the designer had to hit a comparatively infinitesimal target in design space just to create a working car, let alone a car of the Corolla's quality. You cannot build so much as an effective wagon by sawing boards randomly and nailing according to coinflips. To hit such a tiny target in configuration space requires a powerful optimization process.

The notion of an "optimization process" is *predictively useful* because it can be easier to understand the *target* of an optimization process than to understand its step-by-step *dynamics*. The above discussion of the Corolla assumes *implicitly* that the designer of the Corolla was trying to produce a "vehicle," a means of travel. This assumption deserves to be made explicit, but it is not wrong, and it is highly useful in understanding the Corolla.

5.2. Aiming at the Target

The temptation is to ask what "AIs" will "want," forgetting that the space of minds-in-general is much wider than the tiny human dot. One should resist the temptation to spread quantifiers over all possible minds. Storytellers spinning tales of the distant and exotic land called Future, say how the future *will be*. They make *predictions*. They say, "AIs will attack humans with marching robot armies" or "AIs will invent a cure for cancer." They do not propose complex relations between initial conditions and outcomes—that would lose the audience. But we need relational understanding to *manipulate* the

future, steer it into a region palatable to humankind. If we do not steer, we run the danger of ending up where we are going.

The critical challenge is not to predict that "AIs" will attack humanity with marching robot armies, or alternatively invent a cure for cancer. The task is not even to make the prediction for an *arbitrary* individual AI design. Rather the task is choosing into existence some *particular* powerful optimization process whose beneficial effects can legitimately be asserted.

I *strongly urge* my readers not to start thinking up reasons why a fully generic optimization process would be friendly. Natural selection isn't friendly, nor does it hate you, nor will it leave you alone. Evolution cannot be so anthropomorphized, it does not work like you do. Many pre-1960s biologists expected natural selection to do all sorts of nice things, and rationalized all sorts of elaborate reasons why natural selection would do it. They were disappointed, because natural selection itself did not start out knowing that it wanted a humanly-nice result, and then rationalize elaborate ways to produce nice results using selection pressures. Thus the events in Nature were outputs of causally different process from what went on in the pre-1960s biologists' minds, so that prediction and reality diverged.

Wishful thinking adds detail, constrains prediction, and thereby creates a burden of improbability. What of the civil engineer who hopes a bridge won't fall? Should the engineer argue that bridges in general are not likely to fall? But Nature itself does not rationalize reasons why bridges should not fall. Rather the civil engineer overcomes the burden of improbability through specific choice guided by specific understanding. A civil engineer starts by desiring a bridge; then uses a rigorous theory to select a bridge design which supports cars; then builds a real-world bridge whose structure reflects the calculated design; and thus the real-world structure supports cars. Thus achieving harmony of predicted positive results and actual positive results.

6. Friendly AI

It would be a very good thing if humanity knew how to choose into existence a powerful optimization process with a particular target. Or in more colloquial terms, it would be nice if we knew how to build a nice AI.

To describe the *field of knowledge* needed to address that challenge, I have proposed the term "Friendly AI." In addition to referring to a body of technique, "Friendly AI" might also refer to the *product* of technique—an AI created with specified motivations. When I use the term *Friendly* in either sense, I capitalize it to avoid confusion with the intuitive sense of "friendly."

One common reaction I encounter is for people to immediately declare that Friendly AI is an impossibility, because any sufficiently powerful AI will be able to modify its own source code to break any constraints placed upon it.

The first flaw you should notice is a Giant Cheesecake Fallacy. Any AI with free access to its own source would, in principle, possess the *ability* to modify its own source code in a way that changed the AI's optimization target. This does not imply the AI has the *motive* to change its own motives. I would not knowingly swallow a pill that made me enjoy committing murder, because *currently* I prefer that my fellow humans not die.

But what if I try to modify myself, and make a mistake? When computer engineers prove a chip valid—a good idea if the chip has 155 million transistors and you can't issue a patch afterward—the engineers use human-guided, machine-verified formal proof. The glorious thing about *formal* mathematical proof, is that a proof of ten billion steps is just as reliable as a proof of ten steps. But human beings are not trustworthy to peer over a purported proof of ten billion steps; we have too high a chance of missing an error. And present-day theorem-proving techniques are not smart enough to design and prove an entire computer chip on their own—current algorithms undergo an exponential explosion in the search space. Human mathematicians can prove theorems far more complex than modern theorem-provers can handle, without being defeated by exponential explosion. But human mathematics is informal and unreliable; occasionally someone discovers a flaw in a previously accepted informal proof. The upshot is that human engineers guide a theorem-prover through the *intermediate* steps of a proof. The human chooses the next lemma, and a complex theorem-prover generates a formal proof, and a simple verifier checks the steps. That's how modern engineers build reliable machinery with 155 million interdependent parts.

Proving a computer chip correct requires a synergy of human intelligence and computer algorithms, as *currently* neither suffices on its own. Perhaps a true AI could use a similar *combination of abilities* when modifying its own code—would have *both* the capability to *invent* large designs without being defeated by exponential explosion, and *also* the ability to *verify* its steps with extreme reliability. That is one way a true AI might remain knowably stable in its goals, even after carrying out a large number of self-modifications.

This paper will not explore the above idea in detail. (Though see Schmidhuber (2007) for a related notion.) But one ought to think about a challenge, and study it in the best available technical detail, *before* declaring it impossible—especially if great stakes depend upon the answer. It is disrespectful to human ingenuity to declare a challenge unsolvable without taking a close look and exercising creativity. It is an enormously strong statement to say that you *cannot* do a thing—that you *cannot* build a heavier-than-air flying machine, that you *cannot* get useful energy from nuclear reactions, that you *cannot* fly to

the Moon. Such statements are universal generalizations, quantified over every single approach that anyone ever has or ever will think up for solving the problem. It only takes a single counterexample to falsify a universal quantifier. The statement that Friendly (or friendly) AI is *theoretically impossible*, dares to quantify over *every possible* mind design and *every possible* optimization process—including human beings, who are also minds, some of whom are nice and wish they were nicer. At this point there are any number of vaguely plausible reasons why Friendly AI might be *humanly* impossible, and it is still more likely that the problem is solvable but no one will get around to solving it in time. But one should not so quickly write off the challenge, especially considering the stakes.

7. Technical Failure and Philosophical Failure

Bostrom (2002) defines an existential catastrophe as one which permanently extinguishes Earth-originating intelligent life *or destroys a part of its potential.* We can divide potential failures of attempted Friendly AI into two informal fuzzy categories, *technical failure* and *philosophical failure*. Technical failure is when you try to build an AI and it doesn't work the way you think it does—you have failed to understand the true workings of your own code. Philosophical failure is trying to build the wrong thing, so that even if you succeeded you would still fail to help anyone or benefit humanity. Needless to say, the two failures are not mutually exclusive.

The border between these two cases is thin, since most philosophical failures are much easier to explain in the presence of technical knowledge. In theory you ought first to say what you *want*, then figure out *how* to get it. In practice it often takes a deep technical understanding to figure out what you want.

7.1. An Example of Philosophical Failure

In the late 19th century, many honest and intelligent people advocated communism, all in the best of good intentions. The people who first invented and spread and swallowed the communist meme were, in sober historical fact, idealists. The *first* communists did not have the example of Soviet Russia to warn them. *At the time, without benefit of hindsight, it must have sounded like a pretty good idea.* After the revolution, when communists came into power and were corrupted by it, other motives may have come into play; but this itself was not something the first idealists predicted, however predictable it may have been. It is *important* to understand that the authors of huge catastrophes need not be evil, nor even *unusually* stupid. If we attribute every tragedy to evil or unusual stupidity, we will look at ourselves, correctly perceive that we are not evil or unusually stupid, and say: "But that would never happen to *us*."

What the first communist revolutionaries thought would happen, as the empirical consequence of their revolution, was that people's lives would improve: laborers would no longer work long hours at backbreaking labor and make little money from it. This turned out not to be the case, to put it mildly. But what the first communists *thought* would happen, was not so very different from what advocates of other political systems thought would be the empirical consequence of *their* favorite political systems. They thought people would be happy. They were wrong.

Now imagine that someone should attempt to program a "Friendly" AI to implement communism, or libertarianism, or anarcho-feudalism, or *favoritepoliticalsystem*, believing that this shall bring about utopia. People's favorite political systems inspire blazing suns of positive affect, so the proposal will sound like a really good idea to the proposer.

We could view the programmer's failure on a moral or ethical level—say that it is the result of someone trusting themselves too highly, failing to take into account their own fallibility, refusing to consider the possibility that communism might be mistaken after all. But in the language of Bayesian decision theory, there's a complementary technical view of the problem. From the perspective of decision theory, the choice for communism stems from combining an empirical belief with a value judgment. The *empirical* belief is that communism, when implemented, results in a specific outcome or class of outcomes: people will be happier, work fewer hours, and possess greater material wealth. This is ultimately an *empirical* prediction; even the part about happiness is a real property of brain states, though hard to measure. If you implement communism, either this outcome eventuates or it does not. The value judgment is that this outcome satisfices or is preferable to current conditions. Given a different *empirical* belief about the *actual real-world consequences* of a communist system, the decision may undergo a corresponding change.

We would expect a true AI, an Artificial General Intelligence, to be capable of changing its empirical beliefs (or its probabilistic world-model, et cetera). If somehow Charles Babbage had lived before Nicolaus Copernicus, and somehow computers had been invented before telescopes, and somehow the programmers of that day and age successfully created an Artificial General Intelligence, it would not follow that the AI would believe forever after that the Sun orbited the Earth. The AI might transcend the factual error of its programmers, provided that the programmers understood inference rather better than they understood astronomy. To build an AI that *discovers* the orbits of the planets, the programmers need not know the math of Newtonian mechanics, only the math of Bayesian probability theory.

The folly of programming an AI to implement communism, or any other political system, is that you're programming *means* instead of *ends*. You're programming in a fixed decision, without that decision being re-evaluable after acquiring improved empirical

knowledge about the results of communism. You are giving the AI a fixed decision without telling the AI how to re-evaluate, at a higher level of intelligence, the fallible process which produced that decision.

If I play chess against a stronger player, I cannot predict *exactly* where my opponent will move against me—if I could predict that, I would necessarily be at least that strong at chess myself. But I can predict the end result, which is a win for the other player. I know the region of possible futures my opponent is aiming for, which is what lets me predict the destination, even if I cannot see the path. When I am at my most creative, that is when it is hardest to predict my actions, and *easiest* to predict the *consequences* of my actions. (Providing that you know and understand my goals!) If I want a better-than-human chess player, I have to program a *search* for winning moves. I can't program in specific moves because then the chess player won't be any better than I am. When I launch a search, I necessarily sacrifice my ability to predict the *exact* answer in advance. To get a really good answer you must sacrifice your ability to predict the answer, albeit not your ability to say what is the question.

Such confusion as to program in communism directly, probably would not tempt an AGI programmer who speaks the language of decision theory. I would call it a philosophical failure, but blame it on lack of technical knowledge.

7.2. An Example of Technical Failure

> In place of laws constraining the behavior of intelligent machines, we need to give them emotions that can guide their learning of behaviors. They should want us to be happy and prosper, which is the emotion we call love. We can design intelligent machines so their primary, innate emotion is unconditional love for all humans. First we can build relatively simple machines that learn to recognize happiness and unhappiness in human facial expressions, human voices and human body language. Then we can hard-wire the result of this learning as the innate emotional values of more complex intelligent machines, positively reinforced when we are happy and negatively reinforced when we are unhappy. Machines can learn algorithms for approximately predicting the future, as for example investors currently use learning machines to predict future security prices. So we can program intelligent machines to learn algorithms for predicting future human happiness, and use those predictions as emotional values. (Hibbard 2001)

Once upon a time, the US Army wanted to use neural networks to automatically detect camouflaged enemy tanks. The researchers trained a neural net on 50 photos of camouflaged tanks in trees, and 50 photos of trees without tanks. Using standard techniques for supervised learning, the researchers trained the neural network to a weighting that

correctly loaded the training set—output "yes" for the 50 photos of camouflaged tanks, and output "no" for the 50 photos of forest. This did not ensure, or even imply, that *new* examples would be classified correctly. The neural network might have "learned" 100 special cases that would not generalize to any new problem. Wisely, the researchers had originally taken 200 photos, 100 photos of tanks and 100 photos of trees. They had used only 50 of each for the training set. The researchers ran the neural network on the remaining 100 photos, and without further training the neural network classified all remaining photos correctly. Success confirmed! The researchers handed the finished work to the Pentagon, which soon handed it back, complaining that in their own tests the neural network did no better than chance at discriminating photos.

It turned out that in the researchers' dataset, photos of camouflaged tanks had been taken on cloudy days, while photos of plain forest had been taken on sunny days. The neural network had learned to distinguish cloudy days from sunny days, instead of distinguishing camouflaged tanks from empty forest.

A technical failure occurs when the code does not do what you think it does, though it faithfully executes as you programmed it. More than one model can load the same data. Suppose we trained a neural network to recognize smiling human faces and distinguish them from frowning human faces. Would the network classify a tiny picture of a smiley-face into the same attractor as a smiling human face? If an AI "hard-wired" to such code possessed the power—and Hibbard (2001) spoke of superintelligence—would the galaxy end up tiled with tiny molecular pictures of smiley-faces?

This form of failure is especially dangerous because it will *appear* to work within a fixed context, then fail when the context changes. The researchers of the "tank classifier" story tweaked their neural network until it correctly loaded the training data, then verified the network on additional data(without further tweaking). Unfortunately, both the training data and verification data turned out to share an assumption which held over the all data used in development, but not in all the real-world contexts where the neural network was called upon to function. In the story of the tank classifier, the assumption is that tanks are photographed on cloudy days.

Suppose we wish to develop an AI of increasing power. The AI possesses a developmental stage where the human programmers are more powerful than the AI—not in the sense of mere physical control over the AI's electrical supply, but in the sense that the human programmers are smarter, more creative, more cunning than the AI. During the developmental period we suppose that the programmers possess the ability to make changes to the AI's source code without needing the consent of the AI to do so. However, the AI is also intended to possess postdevelopmental stages, including, in the case of Hibbard's scenario, superhuman intelligence. An AI of superhuman intelligence surely could not be modified without its consent. At this point we must rely on the pre-

viously laid-down goal system to function correctly, because if it operates in a sufficiently unforeseen fashion, the AI may actively resist our attempts to correct it—and, if the AI is smarter than a human, probably win.

Trying to control a growing AI by *training a neural network to provide its goal system* faces the problem of a huge *context change* between the AI's developmental stage and postdevelopmental stage. During the developmental stage, the AI may *only* be able to produce stimuli that fall into the "smiling human faces" category, by solving humanly provided tasks, as its makers intended. Flash forward to a time when the AI is superhumanly intelligent and has built its own nanotech infrastructure, and the AI may be able to produce stimuli classified into the same attractor by tiling the galaxy with tiny smiling faces.

Thus the AI appears to work fine during development, but produces catastrophic results after it becomes smarter than the programmers(!).

There is a temptation to think, "But surely the AI will know that's not what we meant?" But the code is not *given* to the AI, for the AI to look over and hand back if it does the wrong thing. The code *is* the AI. Perhaps with enough effort and understanding we can write code that cares if we have written the wrong code—the legendary DWIM instruction, which among programmers stands for Do-What-I-Mean (Raymond 2003). But effort is required to write a DWIM dynamic, and nowhere in Hibbard's proposal is there mention of designing an AI that does what we mean, not what we say. Modern chips don't DWIM their code; it is not an automatic property. And if you messed up the DWIM itself, you would suffer the consequences. For example, suppose DWIM was defined as maximizing the satisfaction of the programmer with the code; when the code executed as a superintelligence, it might rewrite the programmers' brains to be maximally satisfied with the code. I do not say this is inevitable; I only point out that Do-What-I-Mean is a major, nontrivial technical challenge of Friendly AI.

8. Rates of Intelligence Increase

From the standpoint of existential risk, one of the most critical points about Artificial Intelligence is that an Artificial Intelligence might increase in intelligence *extremely fast*. The obvious reason to suspect this possibility is recursive self-improvement (Good 1965). The AI becomes smarter, including becoming smarter at the task of writing the internal cognitive functions of an AI, so the AI can rewrite its existing cognitive functions to work even better, which makes the AI still smarter, including smarter at the task of rewriting itself, so that it makes yet more improvements.

Human beings do not recursively self-improve in a *strong* sense. To a *limited* extent, we improve ourselves: we learn, we practice, we hone our skills and knowledge.

To a *limited* extent, these self-improvements improve our ability to improve. New discoveries can increase our ability to make further discoveries—in that sense, knowledge feeds on itself. But there is still an underlying level we haven't yet touched. We haven't rewritten the human brain. The brain is, ultimately, the source of discovery, and our brains today are much the same as they were ten thousand years ago.

In a similar sense, natural selection improves organisms, but the process of natural selection does not itself improve—not in a strong sense. Adaptation can open up the way for additional adaptations. In this sense, adaptation feeds on itself. But even as the gene pool boils, there's still an underlying heater, the process of mutation and recombination and selection, which is not itself re-architected. A few rare innovations increased the rate of evolution itself, such as the invention of sexual recombination. But even sex did not change the essential nature of evolution: its lack of abstract intelligence, its reliance on random mutations, its blindness and incrementalism, its focus on allele frequencies. Similarly, not even the invention of science changed the essential character of the human brain: its limbic core, its cerebral cortex, its prefrontal self-models, its characteristic speed of 200 Hz.

An Artificial Intelligence could rewrite its code from scratch—it could change the underlying dynamics of optimization. Such an optimization process would wrap around *much more strongly* than either evolution accumulating adaptations, or humans accumulating knowledge. The key implication for our purposes is that an AI might make a *huge* jump in intelligence after reaching some threshold of criticality.

One often encounters skepticism about this scenario—what Good (1965) called an "intelligence explosion"—because progress in Artificial Intelligence has the reputation of being very slow. At this point it may prove helpful to review a loosely analogous historical surprise. (What follows is taken primarily from Rhodes [1986].)

In 1933, Lord Ernest Rutherford said that no one could ever expect to derive power from splitting the atom: "Anyone who looked for a source of power in the transformation of atoms was talking moonshine." At that time laborious hours and weeks were required to fission a handful of nuclei.

Flash forward to 1942, in a squash court beneath Stagg Field at the University of Chicago. Physicists are building a shape like a giant doorknob out of alternate layers of graphite and uranium, intended to start the first self-sustaining nuclear reaction. In charge of the project is Enrico Fermi. The key number for the pile is k, the effective neutron multiplication factor: the average number of neutrons from a fission reaction that cause another fission reaction. At k less than one, the pile is subcritical. At $k \geq 1$, the pile should sustain a critical reaction. Fermi calculates that the pile will reach $k = 1$ between layers 56 and 57.

A work crew led by Herbert Anderson finishes Layer 57 on the night of December 1, 1942. Control rods, strips of wood covered with neutron-absorbing cadmium foil, prevent the pile from reaching criticality. Anderson removes all but one control rod and measures the pile's radiation, confirming that the pile is ready to chain-react the next day. Anderson inserts all cadmium rods and locks them into place with padlocks, then closes up the squash court and goes home.

The next day, December 2, 1942, on a windy Chicago morning of sub-zero temperatures, Fermi begins the final experiment. All but one of the control rods are withdrawn. At 10:37am, Fermi orders the final control rod withdrawn about half-way out. The geiger counters click faster, and a graph pen moves upward. "This is not it," says Fermi, "the trace will go to this point and level off," indicating a spot on the graph. In a few minutes the graph pen comes to the indicated point, and does not go above it. Seven minutes later, Fermi orders the rod pulled out another foot. Again the radiation rises, then levels off. The rod is pulled out another six inches, then another, then another. At 11:30, the slow rise of the graph pen is punctuated by an enormous CRASH—an emergency control rod, triggered by an ionization chamber, activates and shuts down the pile, which is still short of criticality. Fermi calmly orders the team to break for lunch.

At 2pm the team reconvenes, withdraws and locks the emergency control rod, and moves the control rod to its last setting. Fermi makes some measurements and calculations, then again begins the process of withdrawing the rod in slow increments. At 3:25pm, Fermi orders the rod withdrawn another twelve inches. "This is going to do it," Fermi says. "Now it will become self-sustaining. The trace will climb and continue to climb. It will not level off."

Herbert Anderson recounts (440):

> At first you could hear the sound of the neutron counter, clickety-clack, clickety-clack. Then the clicks came more and more rapidly, and after awhile they began to merge into a roar; the counter couldn't follow anymore. That was the moment to switch to the chart recorder. But when the switch was made, everyone watched in the sudden silence the mounting deflection of the recorder's pen. It was an awesome silence. Everyone realized the significance of that switch; we were in the high intensity regime and the counters were unable to cope with the situation anymore. Again and again, the scale of the recorder had to be changed to accommodate the neutron intensity which was increasing more and more rapidly. Suddenly Fermi raised his hand. "The pile has gone critical," he announced. No one present had any doubt about it.

Fermi kept the pile running for twenty-eight minutes, with the neutron intensity doubling every two minutes. The first critical reaction had k of 1.0006. Even at $k = 1.0006$, the pile was only controllable because some of the neutrons from a uranium fission reac-

tion are *delayed*—they come from the decay of short-lived fission byproducts. For every 100 fissions in U_{235}, 242 neutrons are emitted almost immediately (0.0001s), and 1.58 neutrons are emitted an average of ten seconds later. Thus the *average* lifetime of a neutron is ~0.1 seconds, implying 1,200 generations in two minutes, and a doubling time of two minutes because 1.0006 to the power of 1,200 is ~2. A nuclear reaction which is *prompt critical* is critical without the contribution of delayed neutrons. If Fermi's pile had been prompt critical with $k = 1.0006$, neutron intensity would have doubled every *tenth* of a second.

The first moral is that confusing the speed of *AI research* with the speed of *a real AI once built* is like confusing the speed of physics research with the speed of nuclear reactions. It mixes up the map with the territory. It took years to get that first pile built, by a small group of physicists who didn't generate much in the way of press releases. But, once the pile was built, interesting things happened on the timescale of nuclear interactions, not the timescale of human discourse. In the nuclear domain, elementary interactions happen much faster than human neurons fire. Much the same may be said of transistors.

Another moral is that there's a huge difference between one self-improvement triggering 0.9994 further improvements on average, and one self-improvement triggering 1.0006 further improvements on average. The nuclear pile didn't cross the critical threshold as the result of the physicists suddenly piling on a lot more material. The physicists piled on material slowly and steadily. Even if there is a smooth underlying curve of brain intelligence as a function of optimization pressure previously exerted on that brain, the curve of *recursive self-improvement* may show a huge leap.

There are also other reasons why an AI might show a sudden huge leap in intelligence. The species *Homo sapiens* showed a sharp jump in the effectiveness of intelligence, as the result of natural selection exerting a more-or-less steady optimization pressure on hominids for millions of years, gradually expanding the brain and prefrontal cortex, tweaking the software architecture. A few tens of thousands of years ago, hominid intelligence crossed some key threshold and made a *huge* leap in real-world effectiveness; we went from savanna to skyscrapers in the blink of an evolutionary eye. This happened with a continuous underlying selection pressure—there wasn't a huge jump in the optimization power of *evolution* when humans came along. The underlying brain architecture was also continuous—our cranial capacity didn't suddenly increase by two orders of magnitude. So it might be that, even if the AI is being elaborated from outside by human programmers, the curve for *effective* intelligence will jump sharply.

Or perhaps someone builds an AI prototype that shows some promising results, and the demo attracts another $100 million in venture capital, and this money purchases a thousand times as much supercomputing power. I doubt a thousandfold in-

crease in hardware would purchase anything like a thousandfold increase in effective intelligence—but mere doubt is not reliable in the absence of any ability to perform an analytical calculation. Compared to chimps, humans have a threefold advantage in brain and a sixfold advantage in prefrontal cortex, which suggests (a) software is more important than hardware and (b) small increases in hardware can support large improvements in software. It is one more point to consider.

Finally, AI may make an *apparently* sharp jump in intelligence purely as the result of anthropomorphism, the human tendency to think of "village idiot" and "Einstein" as the extreme ends of the intelligence scale, instead of nearly indistinguishable points on the scale of minds-in-general. Everything dumber than a dumb human may appear to us as simply "dumb." One imagines the "AI arrow" creeping steadily up the scale of intelligence, moving past mice and chimpanzees, with AIs still remaining "dumb" because AIs can't speak fluent language or write science papers, and then the AI arrow crosses the tiny gap from infra-idiot to ultra-Einstein in the course of one month or some similarly short period. I don't think this *exact* scenario is plausible, mostly because I don't expect the curve of recursive self-improvement to move at a linear creep. But I am not the first to point out that "AI" is a moving target. As soon as a milestone is actually achieved, it ceases to be "AI." This can only encourage procrastination.

Let us concede for the sake of argument that, for all we know (and it seems to me also probable in the real world) that an AI has the capability to make a sudden, sharp, large leap in intelligence. What follows from this?

First and foremost: it follows that a reaction I often hear, "We don't need to worry about Friendly AI because we don't yet have AI," is misguided or downright suicidal. We cannot rely on having distant advance warning before AI is created; past technological revolutions usually did not telegraph themselves to people alive *at the time*, whatever was said afterward in hindsight. The mathematics and techniques of Friendly AI will not materialize from nowhere when needed; it takes years to lay firm foundations. And we need to solve the Friendly AI challenge *before* Artificial General Intelligence is created, not afterward; I shouldn't even have to point this out. There will be difficulties for Friendly AI because the field of AI itself is in a state of low consensus and high entropy. But that doesn't mean we don't need to worry about Friendly AI. It means there will be difficulties. The two statements, sadly, are not remotely equivalent.

The possibility of sharp jumps in intelligence also implies a higher standard for Friendly AI techniques. The technique cannot assume the programmers' ability to monitor the AI *against its will*, rewrite the AI *against its will*, bring to bear the threat of superior military force; nor may the algorithm assume that the programmers control a "reward button" which a smarter AI could wrest from the programmers; et cetera. Indeed no one should be making these assumptions to begin with. The indispensable protection is

an AI that does not *want* to hurt you. Without the indispensable, no auxiliary defense can be regarded as safe. No system is secure that searches for ways to defeat its own security. If the AI would harm humanity in *any* context, you must be doing *something* wrong on a very deep level, laying your foundations awry. You are building a shotgun, pointing the shotgun at your foot, and pulling the trigger. You are deliberately setting into motion a created cognitive dynamic that will seek in some context to hurt you. That is the wrong behavior for the dynamic; write code that does something else instead.

For much the same reason, Friendly AI programmers should assume that the AI has total access to its own source code. If the AI *wants* to modify itself to be no longer Friendly, then Friendliness has *already* failed, at the point when the AI forms that intention. Any solution that relies on the AI not being *able* to modify itself must be broken in some way or other, and will still be broken even if the AI never does modify itself. I do not say it should be the *only* precaution, but the *primary* and *indispensable* precaution is that you choose into existence an AI that does not choose to hurt humanity.

To avoid the Giant Cheesecake Fallacy, we should note that the ability to self-improve does not imply the choice to do so. The *successful* exercise of Friendly AI technique might create an AI which had the *potential* to grow more quickly, but chose instead to grow along a slower and more manageable curve. Even so, after the AI passes the criticality threshold of *potential* recursive self-improvement, you are then operating in a much more dangerous regime. If Friendliness fails, the AI might decide to rush full speed ahead on self-improvement—metaphorically speaking, it would go prompt critical.

I tend to assume arbitrarily large *potential* jumps for intelligence because (a) this is the conservative assumption; (b) it discourages proposals based on building AI without really understanding it; and (c) large potential jumps strike me as probable-in-the-real-world. If I encountered a domain where it was conservative *from a risk-management perspective* to assume slow improvement of the AI, then I would demand that a plan not break down *catastrophically* if an AI lingers at a near-human stage for years or longer. This is not a domain over which I am willing to offer narrow confidence intervals.

9. Hardware

People tend to think of large computers as *the* enabling factor for Artificial Intelligence. This is, to put it mildly, an extremely questionable assumption. Outside futurists discussing Artificial Intelligence talk about hardware progress because hardware progress is easy to measure—in contrast to understanding of intelligence. It is not that there has been no progress, but that the progress cannot be charted on neat PowerPoint graphs. Improvements in understanding are harder to report on, and therefore less reported.

Rather than thinking in terms of the "minimum" hardware "required" for Artificial Intelligence, think of a minimum level of researcher understanding that decreases as a function of hardware improvements. The better the computing hardware, the less understanding you need to build an AI. The extremal case is natural selection, which used a ridiculous amount of brute computational force to create human intelligence using *no* understanding, only nonchance retention of chance mutations.

Increased computing power makes it easier to build AI, but there is no obvious reason why increased computing power would help make the AI Friendly. Increased computing power makes it easier to use brute force; easier to combine poorly understood techniques that work. Moore's Law steadily *lowers* the barrier that keeps us from building AI *without* a deep understanding of cognition.

It is acceptable to fail at AI *and* at Friendly AI. It is acceptable to succeed at AI *and* at Friendly AI. What is not acceptable is succeeding at AI and failing at Friendly AI. Moore's Law makes it easier to do exactly that. "Easier," but thankfully not easy. I doubt that AI will be "easy" at the time it is finally built—simply because there are parties who will exert tremendous effort to build AI, and one of them will succeed after AI first becomes possible to build with tremendous effort.

Moore's Law is an interaction between Friendly AI and other technologies, which adds *oft-overlooked* existential risk to other technologies. We can imagine that molecular nanotechnology is developed by a benign multinational governmental consortium and that they successfully avert the *physical-layer* dangers of nanotechnology. They straightforwardly prevent accidental replicator releases, and with much greater difficulty put global defenses in place against malicious replicators; they restrict access to "root level" nanotechnology while distributing configurable nanoblocks, et cetera (Phoenix and Treder 2008). But nonetheless nanocomputers become widely available, either because attempted restrictions are bypassed, or because no restrictions are attempted. And then someone brute-forces an Artificial Intelligence which is nonFriendly; and so the curtain is rung down. This scenario is especially worrying because incredibly powerful nanocomputers would be among the first, the easiest, and the safest-seeming applications of molecular nanotechnology.

What of regulatory controls on supercomputers? I certainly wouldn't rely on it to prevent AI from ever being developed; yesterday's supercomputer is tomorrow's laptop. The standard reply to a regulatory proposal is that when nanocomputers are outlawed, only outlaws will have nanocomputers. The burden is to argue that the supposed benefits of *reduced* distribution outweigh the inevitable risks of *uneven* distribution. For myself I would certainly not argue in *favor* of regulatory restrictions on the use of supercomputers for Artificial Intelligence research; it is a proposal of dubious benefit which would be fought tooth and nail by the entire AI community. But in the unlikely event

that a proposal made it that far through the political process, I would not expend any significant effort on *fighting* it, because I don't expect the good guys to *need* access to the "supercomputers" of their day. *Friendly* AI is not about brute-forcing the problem.

I can imagine regulations effectively controlling a small set of ultra-expensive computing resources that are *presently considered* "supercomputers." But computers are everywhere. It is not like the problem of nuclear proliferation, where the main emphasis is on controlling plutonium and enriched uranium. The raw materials for AI are *already* everywhere. That cat is so far out of the bag that it's in your wristwatch, cellphone, and dishwasher. This too is a special and unusual factor in Artificial Intelligence as an existential risk. We are separated from the risky regime, not by large visible installations like isotope centrifuges or particle accelerators, but *only* by missing knowledge. To use a perhaps over-dramatic metaphor, imagine if subcritical masses of enriched uranium had powered cars and ships throughout the world, *before* Leo Szilard first thought of the chain reaction.

10. Threats and Promises

It is a risky intellectual endeavor to predict *specifically* how a benevolent AI would help humanity, or an unfriendly AI harm it. There is the risk of *conjunction fallacy*: added detail necessarily reduces the joint probability of the entire story, but subjects often assign higher probabilities to stories which include strictly added details (Yudkowsky 2008). There is the risk—virtually the certainty—of failure of imagination; and the risk of Giant Cheesecake Fallacy that leaps from capability to motive. Nonetheless I will try to solidify threats and promises.

The future has a reputation for accomplishing feats which the past thought impossible. Future civilizations have even broken what past civilizations thought (incorrectly, of course) to be the laws of physics. If prophets of 1900 AD—never mind 1000 AD—had tried to bound the powers of human civilization a billion years later, some of those impossibilities would have been accomplished before the century was out; transmuting lead into gold, for example. Because we remember future civilizations surprising past civilizations, it has become cliche that we can't put limits on our great-grandchildren. And yet everyone in the 20th century, in the 19th century, and in the 11th century, was human.

We can distinguish three families of unreliable metaphors for imagining the capability of a smarter-than-human Artificial Intelligence:

***g*-factor metaphors:** Inspired by differences of individual intelligence between humans. AIs will patent new technologies, publish groundbreaking research papers, make money on the stock market, or lead political power blocs.

History metaphors: Inspired by knowledge differences between past and future human civilizations. AIs will swiftly invent the kind of capabilities that cliche would attribute to human civilization a century or millennium from now: molecular nanotechnology; interstellar travel; computers performing 10^{25} operations per second.

Species metaphors: Inspired by differences of brain architecture between species. AIs have magic.

g-factor metaphors seem most common in popular futurism: when people think of "intelligence" they think of human geniuses instead of humans. In stories about hostile AI, *g* metaphors make for a Bostromian "good story": an opponent that is powerful enough to create dramatic tension, but not powerful enough to instantly squash the heroes like bugs, and ultimately weak enough to lose in the final chapters of the book. Goliath against David is a "good story," but Goliath against a fruit fly is not.

If we suppose the *g*-factor metaphor, then global catastrophic risks of this scenario are relatively mild; a hostile AI is not much more of a threat than a hostile human genius. If we suppose a *multiplicity* of AIs, then we have a metaphor of conflict between nations, between the AI tribe and the human tribe. If the AI tribe wins in military conflict and wipes out the humans, that is an existential catastrophe of the Bang variety (Bostrom 2002). If the AI tribe dominates the world economically and attains effective control of the destiny of Earth-originating intelligent life, but the AI tribe's goals do not seem to us interesting or worthwhile, then that is a Shriek, Whimper, or Crunch.

But how likely is it that Artificial Intelligence will cross all the vast gap from amoeba to village idiot, and then stop at the level of human genius?

The fastest observed neurons fire 1000 times per second; the fastest axon fibers conduct signals at 150 meters/second, a half-millionth the speed of light; each synaptic operation dissipates around 15,000 attojoules, which is more than a million times the thermodynamic minimum for irreversible computations at room temperature ($kT_{300} \ln(2) = 0.003$ attojoules per bit). It would be physically possible to build a brain that computed a million times as fast as a human brain, without shrinking the size, or running at lower temperatures, or invoking reversible computing or quantum computing. If a human mind were thus accelerated, a subjective year of thinking would be accomplished for every 31 physical seconds in the outside world, and a millennium would fly by in eight and a half hours. Vinge (1993) referred to such sped-up minds as "weak superhumanity": a mind that thinks like a human but much faster.

We suppose there comes into existence an extremely fast mind, embedded in the midst of human technological civilization as it exists at that time. The failure of imagination is to say, "No matter how fast it thinks, it can only affect the world at the speed of its manipulators; it can't operate machinery faster than it can order human hands to

work; therefore a fast mind is no great threat." It is no law of Nature that physical operations must crawl at the pace of long seconds. Critical times for elementary molecular interactions are measured in femtoseconds, sometimes picoseconds. Drexler (1992) has analyzed controllable molecular manipulators which would complete $>10^6$ mechanical operations per second—note that this is in keeping with the general theme of "million-fold speedup." (The smallest physically sensible increment of time is generally thought to be the Planck interval, $5 \cdot 10^{-44}$ seconds, on which scale even the dancing quarks are statues.)

Suppose that a human civilization were locked in a box and allowed to affect the outside world only through the glacially slow movement of alien tentacles, or mechanical arms that moved at microns per second. We would focus all our creativity on finding the *shortest possible path* to building fast manipulators in the outside world. Pondering fast manipulators, one immediately thinks of molecular nanotechnology—though there may be other ways. What is the *shortest* path you could take to molecular nanotechnology in the slow outside world, if you had eons to ponder each move? The answer is that I don't know because I don't have eons to ponder. Here's one imaginable fast pathway:

1. Crack the protein folding problem, to the extent of being able to generate DNA strings whose folded peptide sequences fill specific functional roles in a complex chemical interaction.
2. Email sets of DNA strings to one or more online laboratories which offer DNA synthesis, peptide sequencing, and FedEx delivery. (Many labs currently offer this service, and some boast of 72-hour turnaround times.)
3. Find at least one human connected to the Internet who can be paid, blackmailed, or fooled by the right background story, into receiving FedExed vials and mixing them in a specified environment.
4. The synthesized proteins form a very primitive "wet" nanosystem which, ribosome-like, is capable of accepting external instructions; perhaps patterned acoustic vibrations delivered by a speaker attached to the beaker.
5. Use the extremely primitive nanosystem to build more sophisticated systems, which construct still more sophisticated systems, bootstrapping to molecular nanotechnology—or beyond.

The elapsed turnaround time would be, imaginably, on the order of a week from when the fast intelligence first became able to solve the protein folding problem. Of course this whole scenario is strictly something *I* am thinking of. Perhaps in 19,500 years of subjective time (one week of physical time at a millionfold speedup) I would think of a better way. Perhaps you can pay for rush courier delivery instead of FedEx. Perhaps

there are existing technologies, or slight modifications of existing technologies, that combine synergetically with simple protein machinery. Perhaps if you are *sufficiently* smart, you can use waveformed electrical fields to alter reaction pathways in existing biochemical processes. I don't know. I'm not that smart.

The challenge is to chain your capabilities—the physical-world analogue of combining weak vulnerabilities in a computer system to obtain root access. If one path is blocked, you choose another, seeking always to increase your capabilities and use them in synergy. The presumptive goal is to obtain *rapid infrastructure,* means of manipulating the external world on a large scale in fast time. Molecular nanotechnology fits this criterion, first because its elementary operations are fast, and second because there exists a ready supply of precise parts—atoms—which can be used to self-replicate and exponentially grow the nanotechnological infrastructure. The pathway alleged above has the AI obtaining rapid infrastructure within a week—this sounds fast to a human with 200 Hz neurons, but is a vastly longer time for the AI.

Once the AI possesses rapid infrastructure, further events happen on the AI's timescale, not a human timescale (unless the AI *prefers* to act on a human timescale). With molecular nanotechnology, the AI could (potentially) rewrite the solar system unopposed.

An unFriendly AI with molecular nanotechnology (or other rapid infrastructure) need not bother with marching robot armies or blackmail or subtle economic coercion. The unFriendly AI has the ability to repattern all matter in the solar system according to its optimization target. This is fatal for us if the AI does not choose *specifically* according to the criterion of how this transformation affects existing patterns such as biology and people. The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else. The AI runs on a different timescale than you do; by the time your neurons finish thinking the words "I should do something" you have already lost.

A Friendly AI plus molecular nanotechnology is presumptively powerful enough to solve any problem which can be solved either by moving atoms or by creative thinking. One should beware of failures of imagination: Curing cancer is a popular contemporary target of philanthropy, but it does not follow that a Friendly AI with molecular nanotechnology would say to itself, "Now I shall cure cancer." Perhaps a better way to view the problem is that biological cells are not programmable. To solve the latter problem cures cancer as a special case, along with diabetes and obesity. A fast, nice intelligence wielding molecular nanotechnology is power on the order of *getting rid of disease,* not *getting rid of cancer.*

There is finally the family of *species metaphors,* based on between-species differences of intelligence. The AI has magic—not in the sense of incantations and potions, but in

the sense that a wolf cannot understand how a gun works, or what sort of effort goes into making a gun, or the nature of that human power which lets us invent guns. Vinge (1993) wrote:

> Strong superhumanity would be more than cranking up the clock speed on a human-equivalent mind. It's hard to say precisely what strong superhumanity would be like, but the difference appears to be profound. Imagine running a dog mind at very high speed. Would a thousand years of doggy living add up to any human insight?

The species metaphor would seem the nearest analogy *a priori*, but it does not lend itself to making up detailed stories. The main advice the metaphor gives us is that *we had better get Friendly AI right,* which is good advice in any case. The only defense it suggests against hostile AI is *not to build it in the first place,* which is also excellent advice. Absolute power is a conservative engineering assumption in Friendly AI, exposing broken designs. If an AI will hurt you given magic, the Friendliness architecture is wrong.

11. Local and Majoritarian Strategies

One may classify proposed risk-mitigation strategies into:

- Strategies which require *unanimous* cooperation; strategies which can be catastrophically defeated by individual defectors or small groups.
- Strategies which require *majority* action; a majority of a legislature in a single country, or a majority of voters in a country, or a majority of countries in the UN: the strategy requires *most, but not all,* people in a large pre-existing group to behave a particular way.
- Strategies which require *local* action—a concentration of will, talent, and funding which overcomes the threshold of some specific task.

Unanimous strategies are unworkable, which does not stop people from proposing them.

A *majoritarian* strategy is sometimes workable, if you have decades in which to do your work. One must build a movement, from its first beginnings over the years, to its debut as a recognized force in public policy, to its victory over opposing factions. Majoritarian strategies take substantial time and *enormous* effort. People have set out to do such, and history records some successes. But beware: history books tend to focus selectively on movements that have an impact, as opposed to the vast majority that never amount to anything. There is an element involved of luck, and of the public's prior willingness to hear. Critical points in the strategy will involve events beyond your

personal control. If you are not willing to devote your entire life to pushing through a majoritarian strategy, don't bother; and just *one* life devoted won't be enough, either.

Ordinarily, *local* strategies are most plausible. A hundred million dollars of funding is not *easy* to obtain, and a global political change is not *impossible* to push through, but it is still vastly easier to obtain a hundred million dollars of funding than to push through a global political change.

Two assumptions which give rise to a *majoritarian* strategy for AI are these:

- A majority of Friendly AIs can effectively protect the human species from a few unFriendly AIs.
- The first AI built cannot by itself do catastrophic damage.

This reprises essentially the situation of a human civilization before the development of nuclear and biological weapons: most people are cooperators in the overall social structure, and defectors can do damage but not *global catastrophic* damage. Most AI researchers will not want to make unFriendly AIs. So long as *someone* knows how to build a stably Friendly AI—so long as the problem is not completely beyond contemporary knowledge and technique—researchers will learn from each other's successes and repeat them. Legislation could (for example) require researchers to publicly report their Friendliness strategies, or penalize researchers whose AIs cause damage; and while this legislation will not prevent *all* mistakes, it may suffice that a *majority* of AIs are built Friendly.

We can also imagine a scenario that implies an easy local strategy:

- The first AI cannot by itself do catastrophic damage.
- If even a single Friendly AI exists, that AI *plus* human institutions can fend off any number of unFriendly AIs.

The easy scenario would hold if e.g., human institutions can reliably distinguish Friendly AIs from unFriendly, and give revocable power into the hands of Friendly AIs. Thus we could pick and choose our allies. The only requirement is that the Friendly AI problem must be solvable (as opposed to being completely beyond human ability).

Both of the above scenarios assume that the *first* AI (the first powerful, general AI) cannot by itself do global catastrophic damage. Most concrete visualizations which imply this use a *g* metaphor: AIs as analogous to unusually able humans. In Section 8 on *rates of intelligence increase*, I listed some reasons to be wary of a *huge, fast* jump in intelligence:

- The distance from idiot to Einstein, which looms large to us, is a small dot on the scale of minds-in-general.

- Hominids made a *sharp* jump in *real-world effectiveness* of intelligence, despite natural selection exerting roughly steady optimization pressure on the underlying genome.
- An AI may absorb a huge amount of additional hardware after reaching some brink of competence (i.e., eat the Internet).
- Criticality threshold of recursive self-improvement. One self-improvement triggering 1.0006 self-improvements is qualitatively different from one self-improvement triggering 0.9994 self-improvements.

As described in Section 10, a sufficiently powerful intelligence may need only a short time (from a human perspective) to achieve molecular nanotechnology, or some other form of rapid infrastructure.

We can therefore visualize a possible *first-mover effect* in superintelligence. The first-mover effect is when the outcome for Earth-originating intelligent life depends primarily on the makeup of whichever mind *first* achieves some key threshold of intelligence—such as criticality of self-improvement. The two necessary assumptions are these:

- The *first* AI to surpass some key threshold (e.g. criticality of self-improvement), if unFriendly, can wipe out the human species.
- The *first* AI to surpass the same threshold, if Friendly, can prevent a hostile AI from coming into existence or from harming the human species; or find some other creative way to ensure the survival and prosperity of Earth-originating intelligent life.

More than one scenario qualifies as a first-mover effect. Each of these examples reflects a different key threshold:

- Post-criticality, self-improvement reaches superintelligence on a timescale of weeks or less. AI projects are sufficiently sparse that no *other* AI achieves criticality before the *first* mover is powerful enough to overcome all opposition. The key threshold is criticality of recursive self-improvement.
- AI-1 cracks protein folding three days before AI-2. AI-1 achieves nanotechnology six hours before AI-2. With rapid manipulators, AI-1 can (potentially) disable AI-2's R&D before fruition. The runners are close, but whoever crosses the finish line first, wins. The key threshold is rapid infrastructure.
- The first AI to absorb the Internet can (potentially) keep it out of the hands of other AIs. Afterward, by economic domination or covert action or blackmail or supreme ability at social manipulation, the first AI halts or slows other AI projects so that no other AI catches up. The key threshold is absorption of a unique resource.

The human species, *Homo sapiens,* is a first mover. From an evolutionary perspective, our cousins, the chimpanzees, are only a hairbreadth away from us. *Homo sapiens* still wound up with all the technological marbles because we got there a little earlier. Evolutionary biologists are still trying to unravel which order the key thresholds came in, because the first-mover species was first to cross so *many:* Speech, technology, abstract thought. . . . We're still trying to reconstruct which dominos knocked over which other dominos. The upshot is that *Homo sapiens* is first mover beyond the shadow of a contender.

A first-mover effect implies a theoretically localizable strategy (a task that can, in principle, be carried out by a strictly local effort), but it invokes a technical challenge of extreme difficulty. We only need to get Friendly AI right in one place and one time, not every time everywhere. But someone must get Friendly AI right on the first try, *before* anyone else builds AI to a lower standard.

I cannot perform a precise calculation using a precisely confirmed theory, but my *current opinion* is that sharp jumps in intelligence are *possible, likely,* and *constitute the dominant probability.* This is not a domain in which I am willing to give narrow confidence intervals, and therefore a strategy must not fail *catastrophically*—should not leave us worse off than before—if a sharp jump in intelligence does *not* materialize. But a much more serious problem is strategies visualized for slow-growing AIs, which fail catastrophically if there *is* a first-mover effect. This is a more serious problem because:

- Faster-growing AIs represent a greater technical challenge.
- Like a car driving over a bridge built for trucks, an AI designed to remain Friendly in extreme conditions should (presumptively) remain Friendly in less extreme conditions. The reverse is not true.
- Rapid jumps in intelligence are counterintuitive in everyday social reality. The *g*-factor metaphor for AI is intuitive, appealing, reassuring, and conveniently implies fewer design constraints.
- It is my current guess that the curve of intelligence increase *does* contain huge, sharp (potential) jumps.

My current strategic outlook tends to focus on the difficult local scenario: The first AI must be Friendly. With the caveat that, if no sharp jumps in intelligence materialize, it should be possible to switch to a strategy for making a majority of AIs Friendly. In either case, the technical effort that went into preparing for the extreme case of a first mover should leave us better off, not worse.

The scenario that implies an impossible, unanimous strategy is:

- A single AI can be powerful enough to destroy humanity, even despite the protective efforts of Friendly AIs.

- No AI is powerful enough to prevent human researchers from building one AI after another (or find some other creative way of solving the problem).

It is good that this balance of abilities seems unlikely a priori, because in this scenario we are doomed. If you deal out cards from a deck, one after another, you will eventually deal out the ace of clubs.

The same problem applies to the strategy of *deliberately* building AIs that choose not to increase their capabilities past a fixed point. If capped AIs are not powerful enough to defeat uncapped AIs, or prevent uncapped AIs from coming into existence, then capped AIs cancel out of the equation. We keep dealing through the deck until we deal out a superintelligence, whether it is the ace of hearts or the ace of clubs.

A majoritarian strategy only works if it is not *possible* for a single defector to cause global catastrophic damage. For AI, this possibility or impossibility is a natural feature of the design space—the *possibility* is not subject to human decision any more than the speed of light or the gravitational constant.

12. AI Versus Human Intelligence Enhancement

I do not think it plausible that *Homo sapiens* will continue into the indefinite future, thousands or millions of billions of years, without *any* mind *ever* coming into existence that breaks the current upper bound on intelligence. If so, there must come a time when humans *first* face the challenge of smarter-than-human intelligence. If we win the first round of the challenge, then humankind may call upon smarter-than-human intelligence with which to confront later rounds.

Perhaps we would rather take some other route than AI to smarter-than-human intelligence—say, augment humans instead? To pick one extreme example, suppose the one says: The prospect of AI makes me nervous. I would rather that, before any AI is developed, individual humans are scanned into computers, neuron by neuron, and then upgraded, slowly but surely, until they are super-smart; and *that* is the ground on which humanity should confront the challenge of superintelligence.

We are then faced with two questions: Is this scenario possible? And if so, is this scenario desirable? (It is wiser to ask the two questions in that order, for reasons of rationality: we should avoid getting emotionally attached to attractive options that are not actually options.)

Let us suppose an individual human is scanned into a computer, neuron by neuron, as proposed in Moravec (1988). It necessarily follows that the computing capacity used considerably *exceeds* the computing power of the human brain. By hypothesis, the computer runs a detailed simulation of a biological human brain, executed in sufficient fidelity to avoid any detectable high-level effects from systematic low-level errors. Any

accident of biology that affects information-processing *in any way*, we must faithfully simulate to sufficient precision that the overall flow of processing remains isomorphic. To *simulate* the messy biological computer that is a human brain, we need far more *useful* computing power than is embodied in the messy human brain itself.

The most probable way we would develop the ability to scan a human brain neuron by neuron—in sufficient detail to capture *every* cognitively relevant aspect of neural structure—would be the invention of sophisticated molecular nanotechnology. Molecular nanotechnology could probably produce a desktop computer with total processing power exceeding the aggregate brainpower of the entire current human population (Bostrom 1998; Moravec 1999; Merkle and Drexler 1996; Sandberg 1999).

Furthermore, if technology permits us to scan a brain in sufficient fidelity to *execute the scan as code*, it follows that for some years previously, the technology has been available to obtain *extremely detailed* pictures of processing in neural circuitry, and presumably researchers have been doing their best to understand it.

Furthermore, to *upgrade* the upload—transform the brain scan so as to increase the intelligence of the mind within—we must necessarily understand the *high-level* functions of the brain, and how they contribute usefully to intelligence, in excellent detail.

Furthermore, humans are not designed to be improved, either by outside neuroscientists, or by recursive self-improvement internally. Natural selection did not build the human brain to be humanly hackable. All complex machinery in the brain has adapted to operate within narrow parameters of brain design. Suppose you can make the human smarter, let alone superintelligent; does the human remain *sane*? The human brain is very easy to perturb; just changing the balance of neurotransmitters can trigger schizophrenia, or other disorders. Deacon (1997) has an excellent discussion of the evolution of the human brain, how delicately the brain's elements may be balanced, and how this is reflected in modern brain dysfunctions. The human brain is not end-user-modifiable.

All of this makes it rather implausible that the first human being would be *scanned into a computer and sanely upgraded* before *anyone anywhere first built an Artificial Intelligence.*

At the point where technology first becomes capable of uploading, this implies *overwhelmingly more computing power*, and probably *far better cognitive science*, than is required to build an AI.

Building a 747 from scratch is not easy. But is it easier to:

- Start with *the existing design of a biological bird,*
- and *incrementally modify the design through a series of successive stages,*
- each stage *independently viable,*
- such that the endpoint is *a bird scaled up to the size of a 747,*

- which *actually flies,*
- *as fast as a 747,*
- and then *carry out this series of transformations on an actual living bird,*
- *without killing the bird or making it extremely uncomfortable?*

I'm not saying it could never, ever be done. I'm saying that it would be *easier* to build the 747, and then have the 747, metaphorically speaking, upgrade the bird. "Let's just scale up an existing bird to the size of a 747" is *not* a clever strategy that avoids dealing with the intimidating theoretical mysteries of aerodynamics. Perhaps, in the beginning, all you know about flight is that a bird has the mysterious essence of flight, and the materials with which you must build a 747 are just lying there on the ground. But you cannot sculpt the mysterious essence of flight, even as it already resides in the bird, until flight has ceased to be a mysterious essence unto you.

The above argument is directed at a deliberately extreme case. The general point is that we do not have *total* freedom to pick a path that sounds nice and reassuring, or that would make a good story as a science fiction novel. We are constrained by which technologies are likely to precede others.

I am not against scanning human beings into computers and making them smarter, but it seems exceedingly unlikely that this will be the ground on which humanity *first* confronts the challenge of smarter-than-human intelligence. With various *strict subsets* of the technology and knowledge required to *upload and upgrade* humans, one could:

- Upgrade biological brains in-place (for example, by adding new neurons which will be usefully wired in);
- *or* usefully interface computers to biological human brains;
- *or* usefully interface human brains with each other;
- *or* construct Artificial Intelligence.

Furthermore, it is one thing to sanely enhance an average human to IQ 140, and another to enhance a Nobel Prize winner to something beyond human. (Leaving aside quibbles about the suitability of IQ, or Nobel-Prize-winning, as a measure of fluid intelligence; please excuse my metaphors.) Taking Piracetam (or drinking caffeine) may, or may not, make at least some people smarter; but it will not make you *substantially smarter than Einstein.* In which case we haven't won any significant new capabilities; we haven't made further rounds of the problem easier; we haven't broken the upper bound on the intelligence available to deal with existential risks. From the standpoint of managing existential risk, any intelligence enhancement technology which doesn't produce a (nice,

sane) mind literally *smarter than human,* begs the question of whether the same time and effort could be more productively spent to find an extremely smart modern-day human and unleash them on the same problem.

Furthermore, the farther you go from the "natural" design bounds of the human brain—the ancestral condition represented by the brain itself, to which individual brain components are adapted—the greater the danger of individual insanity. If the augment is substantially smarter than human, this too is a global catastrophic risk. How much damage can an evil augmented human do? Well . . . how creative are they? The first question that comes to my mind is, "Creative enough to build their own recursively self-improving AI?"

Radical human intelligence enhancement techniques raise their own safety issues. Again, I am not claiming these problems as engineering impossibilities; only pointing out that the problems exist. AI has safety issues; so does human intelligence enhancement. Not everything that clanks is your enemy, and not everything that squishes is your friend. On the one hand, a nice human *starts out* with all the immense moral, ethical, and architectural complexity that describes what we mean by a "friendly" decision. On the other hand, an AI can be *designed for* stable recursive self-improvement, and shaped to safety: natural selection did not design the human brain with multiple rings of precautionary measures, conservative decision processes, and orders of magnitude of safety margin.

Human intelligence enhancement is a question in its own right, not a subtopic of Artificial Intelligence; and this chapter lacks space to discuss it in detail. It is worth mentioning that I considered both human intelligence enhancement and Artificial Intelligence at the start of my career, and decided to allocate my efforts to Artificial Intelligence. Primarily this was because I did not expect *useful, human-transcending* intelligence enhancement techniques to arrive in time to *substantially impact* the development of recursively self-improving Artificial Intelligence. I would be pleasantly surprised to be proven wrong about this.

But I do not think that it is a viable strategy to deliberately choose *not* to work on Friendly AI, while others work on human intelligence enhancement, in hopes that augmented humans will solve the problem better. I am not willing to embrace a strategy which fails *catastrophically* if human intelligence enhancement takes longer than building AI. (Or vice versa, for that matter.) I fear that working with biology will just take too much time—there will be too much inertia, too much fighting of poor design decisions already made by natural selection. I fear regulatory agencies will not approve human experiments. And even human geniuses take years to learn their art; the faster the augment has to learn, the more difficult it is to augment someone to that level.

I would be pleasantly surprised if augmented humans showed up and built a Friendly AI before anyone else got the chance. But someone who would like to see this outcome will probably have to work hard to speed up intelligence enhancement technologies; it would be difficult to convince me to slow down. If AI is *naturally* far more difficult than intelligence enhancement, no harm done; if building a 747 is *naturally* easier than inflating a bird, then the wait could be fatal. There is a relatively small region of possibility within which deliberately not working on Friendly AI could *possibly* help, and a large region within which it would be either irrelevant or harmful. Even if human intelligence enhancement is possible, there are real, difficult safety considerations; I would have to seriously ask whether we wanted Friendly AI to precede intelligence enhancement, rather than vice versa.

I do not assign strong confidence to the assertion that Friendly AI is easier than human augmentation, or that it is safer. There are many conceivable pathways for augmenting a human. Perhaps there is a technique which is easier and safer than AI, which is also powerful enough to make a difference to existential risk. If so, I may switch jobs. But I did wish to point out some considerations which argue against the *unquestioned assumption* that human intelligence enhancement is easier, safer, and powerful enough to make a difference.

13. Interactions of AI with Other Technologies

Speeding up a desirable technology is a local strategy, while *slowing down* a dangerous technology is a difficult majoritarian strategy. *Halting* or *relinquishing* an undesirable technology tends to require an impossible unanimous strategy. I would suggest that we think, not in terms of developing or not-developing technologies, but in terms of our *pragmatically available latitude* to *accelerate* or *slow down* technologies; and ask, *within the realistic bounds of this latitude,* which technologies we might prefer to see developed *before* or *after* one another.

In nanotechnology, the goal usually presented is to develop defensive shields before offensive technologies. I worry a great deal about this, because a *given level* of offensive technology tends to require much less sophistication than a technology that can defend against it. Offense has outweighed defense during most of civilized history. Guns were developed centuries before bulletproof vests. Smallpox was used as a tool of war before the development of smallpox vaccines. Today there is still no shield that can deflect a nuclear explosion; nations are protected not by defenses that cancel offenses, but by a balance of offensive terror. The nanotechnologists have set themselves an intrinsically difficult problem.

So should we prefer that nanotechnology precede the development of AI, or that AI precede the development of nanotechnology? As presented, this is something of a trick question. The answer has little to do with the intrinsic difficulty of nanotechnology as an existential risk, or the intrinsic difficulty of AI. So far as *ordering* is concerned, the question we should ask is, "Does AI help us deal with nanotechnology? Does nanotechnology help us deal with AI?"

It looks to me like a successful resolution of Artificial Intelligence should help us considerably in dealing with nanotechnology. I cannot see how nanotechnology would make it easier to develop *Friendly* AI. If huge nanocomputers make it easier to develop AI *without* making it easier to solve the particular challenge of Friendliness, that is a *negative* interaction. Thus, all else being equal, I would greatly prefer that Friendly AI *precede* nanotechnology in the *ordering* of technological developments. If we confront the challenge of AI and succeed, we can call on Friendly AI to help us with nanotechnology. If we develop nanotechnology and survive, we still have the challenge of AI to deal with after that.

Generally speaking, a *success* on Friendly AI should help solve nearly any other problem. Thus, if a technology makes AI neither easier nor harder, but carries with it a catastrophic risk, we should prefer all else being equal to *first* confront the challenge of AI.

Any technology which increases available computing power decreases the minimum theoretical sophistication necessary to develop Artificial Intelligence, but doesn't help at all on the *Friendly* side of things, and I count it as a net negative. Moore's Law of Mad Science: Every eighteen months, the minimum IQ necessary to destroy the world drops by one point.

A success on human intelligence enhancement would make Friendly AI easier, and also help on other technologies. But human augmentation is *not* necessarily safer, or easier, than Friendly AI; nor does it necessarily lie within our realistically available latitude to reverse the natural ordering of human augmentation and Friendly AI, if one technology is naturally much easier than the other.

14. Making Progress on Friendly AI

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve

> themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. (McCarthy et al. 1955)

The *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* is the first recorded use of the phrase "artificial intelligence." They had no prior experience to warn them the problem was hard. I would still label it a genuine mistake, that they said "a significant advance *can* be made," not *might* be made, with a summer's work. That is a specific guess about the problem difficulty and solution time, which carries a specific burden of improbability. But if they had said *might*, I would have no objection. How were they to know?

The *Dartmouth Proposal* included, among others, the following topics: Linguistic communication, linguistic reasoning, neural nets, abstraction, randomness and creativity, interacting with the environment, modeling the brain, originality, prediction, invention, discovery, and self-improvement.

Now it seems to me that an AI capable of language, abstract thought, creativity, environmental interaction, originality, prediction, invention, discovery, and above all self-improvement, is *well beyond* the point where it needs also to be Friendly.

The *Dartmouth Proposal* makes no mention of building nice/good/benevolent AI. Questions of safety are not mentioned even for the purpose of dismissing them. This, even in that bright summer when human-level AI seemed just around the corner. The *Dartmouth Proposal* was written in 1955, before the Asilomar conference on biotechnology, thalidomide babies, Chernobyl, or September 11th. If today the idea of artificial intelligence were proposed *for the first time*, then *someone* would demand to know what specifically was being done to manage the risks. I am not saying whether this is a good change or a bad change in our culture. I am not saying whether this produces good or bad science. But the point remains that if the *Dartmouth Proposal* had been written fifty years later, one of the topics would have been safety.

At the time of this writing in 2006, the AI research community still doesn't see Friendly AI as part of the problem. I wish I could cite a reference to this effect, but I cannot cite an a bsence of literature. Friendly AI is absent from the *conceptual* landscape, not just unpopular or unfunded. You cannot even call Friendly AI a blank spot on the map, because there is no notion that something is missing. If you've read popular/semitechnical books proposing how to build AI, such as *Gödel, Escher, Bach* (Hofstadter 1979) or *The Society of Mind* (Minsky 1986), you may think back and recall that you did not see Friendly AI discussed as part of the challenge. Neither have I seen Friendly AI discussed in the technical literature as a technical problem. My attempted literature search turned up primarily brief nontechnical papers, unconnected to each other, with no major reference in common except Isaac Asimov's "Three Laws of Robotics" (Asimov 1942).

Given that this is 2006, why aren't more AI researchers talking about safety? I have no privileged access to others' psychology, but I will briefly speculate based on personal discussions.

The field of Artificial Intelligence has adapted to its experiences over the last fifty years: in particular, the pattern of large promises, especially of human-level capabilities, followed by embarrassing public failure. To attribute this embarrassment to "AI" is perhaps unfair; wiser researchers who made no promises did not see their conservatism trumpeted in the newspapers. Still the failed promises come swiftly to mind, both inside and outside the field of AI, when advanced AI is mentioned. The culture of AI research has adapted to this condition: There is a taboo against talking about human-level capabilities. There is a stronger taboo against anyone who appears to be claiming or predicting a capability they have not demonstrated with running code. The perception I have encountered is that *anyone who claims to be researching Friendly AI is implicitly claiming that their AI design is powerful enough that it needs to be Friendly.*

It should be obvious that this is neither logically true, nor practically a good philosophy. If we imagine someone creating an actual, mature AI which is powerful enough that it *needs* to be Friendly, and moreover, as is our desired outcome, this AI *really is Friendly,* then someone must have been working on Friendly AI for years and years. Friendly AI is not a module you can instantly invent at the exact moment when it is first needed, and then bolt on to an existing, polished design which is otherwise completely unchanged.

The field of AI has techniques, such as neural networks and evolutionary programming, which have grown in power with the slow tweaking of decades. But neural networks are opaque—the user has no idea how the neural net is making its decisions—and cannot easily be rendered unopaque; the people who invented and polished neural networks were not thinking about the long-term problems of Friendly AI. Evolutionary programming (EP) is stochastic, and does not precisely preserve the optimization target in the generated code; EP gives you code that does what you ask, most of the time, under the tested circumstances, but the code may also do something else on the side. EP is a powerful, still maturing technique that is *intrinsically* unsuited to the demands of Friendly AI. Friendly AI, as I have proposed it, requires repeated cycles of recursive self-improvement that precisely preserve a stable optimization target.

The most powerful *current* AI techniques, as they were developed and then polished and improved over time, have basic incompatibilities with the requirements of Friendly AI as I currently see them. The Y2K problem—which proved very expensive to fix, though not global-catastrophic—analogously arose from failing to foresee tomorrow's design requirements. The nightmare scenario is that we find ourselves stuck with a catalog of mature, powerful, publicly available AI techniques which combine to yield *non-*

Friendly AI, but which *cannot* be used to build Friendly AI without redoing the last three decades of AI work from scratch.

In the field of AI it is daring enough to openly discuss *human-level AI*, after the field's past experiences with such discussion. There is the temptation to congratulate yourself on daring so much, and then stop. Discussing *transhuman AI* would seem ridiculous and unnecessary, after daring so much already. (But there is no privileged reason why AIs would slowly climb all the way up the scale of intelligence, and then halt forever exactly on the human dot.) Daring to speak of *Friendly AI*, as a precaution against the global catastrophic risk of *transhuman AI*, would be *two* levels up from the level of daring that is just daring enough to be seen as transgressive and courageous.

There is also a pragmatic objection which concedes that Friendly AI is an important problem, but worries that, given our present state of understanding, we simply are not in a position to tackle Friendly AI: If we try to solve the problem *right now*, we'll just fail, or produce anti-science instead of science.

And this objection is worth worrying about. It appears to me that the knowledge is out there—that it is possible to study a sufficiently large body of existing knowledge, and then tackle Friendly AI without smashing face-first into a brick wall—but the knowledge is scattered across *multiple disciplines*: Decision theory *and* evolutionary psychology *and* probability theory *and* evolutionary biology *and* cognitive psychology *and* information theory *and* the field traditionally known as "Artificial Intelligence" . . . There is no curriculum that has already prepared a large pool of existing researchers to make progress on Friendly AI.

The "ten-year rule" for genius, validated across fields ranging from math to music to competitive tennis, states that no one achieves outstanding performance in any field without at least ten years of effort (Hayes 1981). Mozart began composing symphonies at age 4, but they weren't *Mozart* symphonies—it took another 13 years for Mozart to start composing *outstanding* symphonies (Weisberg 1986). My own experience with the learning curve reinforces this worry. If we want people who can make progress on Friendly AI, then they have to start training themselves, full-time, years before they are *urgently needed*.

If tomorrow the Bill and Melinda Gates Foundation allocated a hundred million dollars of grant money for the study of Friendly AI, then a thousand scientists would at once begin to rewrite their grant proposals to make them appear relevant to Friendly AI. But they would not be genuinely *interested* in the problem—witness that they did not show curiosity before someone offered to pay them. While Artificial General Intelligence is unfashionable and Friendly AI is entirely off the radar, we can at least assume that anyone speaking about the problem is genuinely interested in it. If you throw too

much money at a problem that a field is not prepared to solve, the excess money is more likely to produce anti-science than science—a mess of false solutions.

I cannot regard this verdict as good news. We would all be much safer if Friendly AI could be solved by piling on warm bodies and silver. But as of 2006 I strongly doubt that this is the case—the field of Friendly AI, and Artificial Intelligence itself, is too much in a state of chaos. Yet if the one argues that we *cannot* yet make progress on Friendly AI, that we know too little, we should ask how long the one has studied before coming to this conclusion. Who can say what science does *not* know? There is far too much science for any one human being to learn. Who can say that we are *not* ready for a scientific revolution, in advance of the surprise? And if we *cannot* make progress on Friendly AI because we are not prepared, this does not mean we do not *need* Friendly AI. Those two statements are *not at all* equivalent!

So if we find that we *cannot* make progress on Friendly AI, then we need to figure out how to exit that regime as fast as possible! There is no guarantee whatsoever that, just because we can't manage a risk, the risk will obligingly go away.

If unproven brilliant young scientists become interested in Friendly AI of their own accord, then I think it would be very much to the benefit of the human species if they could apply for a multi-year grant to study the problem full-time. Some funding for Friendly AI is needed to this effect—considerably more funding than presently exists. But I fear that in these beginning stages, a Manhattan Project would only increase the ratio of noise to signal.

15. Conclusion

It once occurred to me that modern civilization occupies an unstable state. I. J. Good's hypothesized intelligence explosion describes a dynamically unstable system, like a pen precariously balanced on its tip. If the pen is *exactly* vertical, it may remain upright; but if the pen tilts even a little from the vertical, gravity pulls it farther in that direction, and the process accelerates. So too would smarter systems have an easier time making themselves smarter.

A dead planet, lifelessly orbiting its star, is also stable. Unlike an intelligence explosion, extinction is not a *dynamic* attractor—there is a large gap between *almost* extinct, and extinct. Even so, *total* extinction is stable.

Must not our civilization eventually wander into one mode or the other?

As logic, the above argument contains holes. Giant Cheesecake Fallacy, for example: minds do not blindly wander into attractors, they have motives. Even so, I suspect that, *pragmatically* speaking, our alternatives boil down to becoming smarter or becoming extinct.

Nature is, not cruel, but indifferent; a neutrality which often seems indistinguishable from outright hostility. Reality throws at you one challenge after another, and when you run into a challenge you can't handle, you suffer the consequences. Often Nature poses requirements that are grossly unfair, even on tests where the penalty for failure is death. How is a 10th-century medieval peasant supposed to invent a cure for tuberculosis? Nature does not match her challenges to your skill, or your resources, or how much free time you have to think about the problem. And when you run into a lethal challenge too difficult for you, you die. It may be unpleasant to think about, but that has been the reality for humans, for thousands upon thousands of years. The same thing could as easily happen to the whole human species, if the human species runs into an unfair challenge.

If human beings did not age, so that 100-year-olds had the same death rate as 15-year-olds, we would not be immortal. We would last only until the probabilities caught up with us. To live even a million years, as an unaging human in a world as risky as our own, you must somehow drive your annual probability of accident down to nearly *zero*. You may not drive; you may not fly; you may not walk across the street even after looking both ways, for it is still too great a risk. Even if you abandoned all thoughts of fun, gave up living to preserve your life, you couldn't navigate a million-year obstacle course. It would be, not physically impossible, but *cognitively* impossible.

The human species, *Homo sapiens*, is unaging but not immortal. Hominids have survived this long only because, for the last million years, there were no arsenals of hydrogen bombs, no spaceships to steer asteroids toward Earth, no biological weapons labs to produce superviruses, no recurring annual prospect of nuclear war or nanotechnological war or rogue Artificial Intelligence. To survive any appreciable time, we need to drive down *each* risk to nearly *zero*. "Fairly good" is not good enough to last another million years.

It seems like an unfair challenge. Such competence is not historically typical of human institutions, no matter how hard they try. For decades the U.S. and the U.S.S.R. avoided nuclear war, but not *perfectly;* there were close calls, such as the Cuban Missile Crisis in 1962. If we postulate that future minds exhibit the same mixture of foolishness and wisdom, the same mixture of heroism and selfishness, as the minds we read about in history books—then the game of existential risk is already over; it was lost from the beginning. We might survive for another decade, even another century, but not another million years.

But the human mind is not the limit of the possible. *Homo sapiens* represents the *first* general intelligence. We were born into the uttermost beginning of things, the dawn of mind. With luck, future historians will look back and describe the present world as an awkward in-between stage of adolescence, when humankind was smart enough to create tremendous problems for itself, but not quite smart enough to solve them.

Yet before we can pass out of that stage of adolescence, we must, as adolescents, confront an adult problem: the challenge of smarter-than-human intelligence. This is the way out of the high-mortality phase of the life cycle, the way to close the window of vulnerability; it is also probably the single most dangerous risk we face. Artificial Intelligence is one road into that challenge; and I think it is the road we will end up taking. I think that, in the end, it will prove easier to build a 747 from scratch, than to scale up an existing bird or graft on jet engines.

I do not want to play down the colossal audacity of trying to build, to a precise purpose and design, something smarter than ourselves. But let us pause and recall that *intelligence* is not the first thing human science has ever encountered which proved difficult to understand. Stars were once mysteries, and chemistry, and biology. Generations of investigators tried and failed to understand those mysteries, and they acquired the reputation of being impossible to mere science. Once upon a time, no one understood why some matter was inert and lifeless, while other matter pulsed with blood and vitality. No one knew how living matter reproduced itself, or why our hands obeyed our mental orders. Lord Kelvin wrote:

> The influence of animal or vegetable life on matter is infinitely beyond the range of any scientific inquiry hitherto entered on. Its power of directing the motions of moving particles, in the demonstrated daily miracle of our human free-will, and in the growth of generation after generation of plants from a single seed, are infinitely different from any possible result of the fortuitous concurrence of atoms. (Macfie 1912)

All scientific ignorance is hallowed by ancientness. Each and every absence of knowledge dates back to the dawn of human curiosity; and the hole lasts through the ages, seemingly eternal, right up until someone fills it. I think it is possible for mere fallible humans to succeed on the challenge of building Friendly AI. But only if intelligence ceases to be a sacred mystery to us, as life was a sacred mystery to Lord Kelvin. Intelligence must cease to be any kind of mystery whatever, sacred or not. We must execute the creation of Artificial Intelligence as the exact application of an exact art. And maybe then we can win.

Artificial Intelligence

According to the father of Artificial Intelligence, John McCarthy, it is "The science and engineering of making intelligent machines, especially intelligent computer programs"

People who are really serious about software should make their own hardware

Author

Dr Prof Engr Mr Santosh Kumar
Senior Technical Officer, Hindustan Aeronautics Limited

Former Software Developer (Microsoft, New York, USA)