

## ***Transcript***

### *Recursive Self-Improvement, and the World's Most Important Math Problem*

Bay Area Future Salon  
February 24, 2006  
Eliezer S. Yudkowsky

## **Introduction**

[Intro slide.] Good evening. I am Eliezer Yudkowsky, Research Fellow of the Singularity Institute for Artificial Intelligence, a 501(c)(3) nonprofit located in Silicon Valley. The title of this talk is "Recursive Self-Improvement and the World's Most Important Math Problem", which was the title I gave the Future Salon when they asked me for a 30-minute talk. Then it turned out I was supposed to give a longer talk, and I got a chance to put in some important parts I thought I'd have to leave out. [Intelligence.] So I'm going to start out by talking about intelligence, but rest assured that I will get to recursive self-improvement and the world's most important math problem. They are still in there, but this way, I get a chance to introduce them properly.

## **Intelligence**

[AI problem.]

The first people who ever used the term Artificial Intelligence defined the term as "making a machine behave in ways that would be called intelligent if a human were so behaving". [McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955.) *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.*] [Turing.] The Turing Test follows essentially the same logic:

- I don't have a good definition of "intelligence".
- However, I know that humans are intelligent.
- If an entity can masquerade as human so well that I can't detect the difference, I will say that this entity is intelligent.

Now, imagine if the Wright Brothers, trying to construct a plane, had reasoned like this: [Bird Test.]

- I don't have a good definition of "flight".
- However, I know that birds can fly.
- If an entity can masquerade as a bird so well that I can't detect the difference, I will say that this entity flies.

Now in point of fact this reasoning works; if you construct something that no observer can tell from a bird, it's got to be able to fly. But you don't have to resort to that kind of reasoning unless flight is a complete mystery to you, so confusing that you can't give a good abstract definition of flight apart from particular things that you think fly.

The first airplane, the Wright Flyer, wasn't built by precisely imitating a bird. The Wright Brothers had the abstract concepts of gravity, lift, and stability. [Wright.] In fact, the Wright Brothers were competent though uncredentialed physicists, who invented a new experimental instrument, the wind tunnel, to test their quantitative predictions. At one point, their predictions failed to match reality, so the Wright Brothers painstakingly tracked down the problem and found an error in an engineering constant, Smeaton's coefficient of air pressure, which had been used for nearly a hundred and fifty years. The Wright Brothers' tale is often told as though two optimistic bicycle mechanics constructed a flying machine despite all the distinguished scientists who said it couldn't be done. Well, there were in fact distinguished scientists who said it couldn't be done. The key difference is not that the Wright Brothers were more optimistic, but that, unlike the distinguished scientists who only spoke their prejudices, the Wright Brothers did the math. Science doesn't work unless you actually use it, no matter how distinguished you are.

[Intelligence.] Okay, so we need some better concept of intelligence before we can build an Artificial Intelligence. When humans try to measure intelligence, we use IQ tests. [IQ]. Which IQ test you use makes surprisingly little difference. [Spearman's g.] Just about any test of general mental ability will correlate with any other one. The correlation component shared out among tests of mental ability is known as Spearman's g; you may have heard of it as g-factor.

The problem with using Spearman's g as a measure of intelligence is that it's

a statistical abstraction from differences of ability *between humans*, who as a *species* are far more intelligent than, say, lizards. Spearman's *g* is abstracted from millimeter height differences among a species of giants. [Fly-Q.] You measure IQ in standard deviations from the human average; there is no absolute zero of IQ that describes a rock. The Wright Brothers would've had a hard time building an airplane, if their concept of flightpower was a Fly-Q test standardized on pigeons.

We should also worry that IQ is too narrow a concept of intelligence, intelligence as equivalent to book smarts. It makes us imagine a starving professor with an IQ of 160 and a billionaire CEO with an IQ of 120. It takes more than book smarts to succeed in the human world: Enthusiasm, social skills, education, musical talent, rationality... but note that each factor I listed is cognitive. It all happens in the brain. And all jokes aside, you won't find many CEOs, nor yet professors of academia, who are chimpanzees.

[Parochial.] When something is universal enough in our everyday lives, we take it for granted to the point of forgetting it exists. When I walked into this room, I didn't ask: Does the room have air? Does the air contain oxygen? I take those everyday things for granted. In the everyday social universe, we take for granted that people are human. We forget the things we have in common, and see only our differences, because only the differences convey new information. We think about the scale of intelligence as if it ran from village idiot up to Einstein. [Cosmopolitan.] What we need is a scale of intelligence that runs from rocks to humans. On that scale, the whole range of human intelligence is pretty much a small dot. Human beings, in general, have a brain three times the size of a chimpanzee brain, and six times as much prefrontal cortex. And that's true whether you're a village idiot or Einstein; there's variance, but not that much variance. A human village idiot has visual cortex, auditory cortex, prefrontal cortex, a thalamus, a hippocampus, an anterior cingulate gyrus and basal ganglia - the whole human architecture. All of this complex machinery underlying intelligence is invisible to us, like the oxygen in the air around us, because in our everyday life it's always there. That's why when people say "intelligence", they think of Einstein, instead of humans.

[Power.] Depending on how you place your markers, the rise of human intelligence in its modern form took place between thirty-five thousand and a hundred and fifty thousand years ago. That event had an enormous,

reshaping effect on the Earth; the land sprouted skyscrapers, planes flew through the skies, footprints appeared on the Moon. If you look around you in this room, most of the things you see are byproducts of human intelligence - the desks, the lights, the floor, all of this comes from the human brain architecture.

Intelligence is power. We forget that sometimes, take our intelligence for granted, just because it's so omnipresent. People say things like "intelligence is no match for a gun", as if guns had grown on trees. People say that intelligence isn't as powerful as money, as if mice used money. People say that intelligence isn't as important as social skills, as if social skills resided in the kidneys instead of the brain. All jokes aside, there are not many famous novelists, or famous military generals, or famous politicians, or famous rationalists, or famous scientists who are chimpanzees. Intelligence is the foundation of human power, the strength that fuels our other arts.

Remember that, when you hear the words "Artificial Intelligence". It's not just chess that we're trying to mess around with, it's the most powerful force in the universe. I say it in all seriousness: intelligence is the most powerful force that human beings have ever tried to mess with. When you ask people for the most powerful force that humans have ever messed with, they'll probably think of nuclear weapons. Well, nuclear weapons didn't grow on trees. They grew on minds.

## **Singularity**

Artificial Intelligence does not belong on the same graph that shows progress in faster aircraft, more fuel-efficient cars, smaller MP3 players, better medical care, et cetera et cetera. You cannot mix Artificial Intelligence into a lumpenfuturism of soaring skyscrapers, healthy cheesecake, and nanotechnological red blood cells that let people hold their breath for sixteen hours. [Unlike the other.] One of these things is not like the other, one of these things doesn't belong. Skyscrapers that get tall enough don't start doing their own engineering. Humanity did not rise to prominence upon Earth by holding its breath longer than other species.

In today's society, most people get their idea of what's important from advertising. And what marketers would have you believe is that important technologies are new, improved, shiny, with extra features, and expensive.

Hollywood, which is part of the same culture, tells you that you're in the Future by showing you soaring skyscrapers and glittering gadgets. And the futurists who make their living as entertainers, who tell fantastic stories to an audience, who are also part of the same culture; they make comfortable and pleasant prophecies that in the future, you will be able to buy ever more cool commodities, maybe with a side order of improved health care. And they treat Artificial Intelligence as a special case of that concept of futurism. But Artificial Intelligence is not just another cool commodity to buy at Fry's Electronics. Our intelligence is the root of all our technology. All our cool gadgets grow out of that. When you mess with intelligence you shake the technology tree by its roots.

Once upon a time there was a professor of mathematics at San Diego State University, who moonlighted as a science-fiction writer. In 1982 this math professor published a brief essay about a very interesting idea. [Lemma.] The first lemma is that if you know exactly what a person smarter than you would do, you would be that smart yourself. For example, if I can predict exactly where Grandmaster Kasparov will move in a chess game, I must be at least that good a chess player myself, because I can just move wherever I think Kasparov would move. And the math professor said: at some point in the future, technology will advance to the point of producing minds which are smarter than human. You can't predict what happens after this point, because if you knew what transhumans would do, you would be that smart yourself. [Misattributed.] The math professor didn't say anything about accelerating change, or Moore's Law, or converging technologies, or emergence, or capabilities going to infinity. There was just this one threshold, creating the first mind that was significantly smarter than human, which would cause a breakdown in the ability of modern humans to predict the future after that point. You could cross the key threshold while technological progress was slowing down, or accelerating - the key threshold was going past human intelligence, and the math professor's observation didn't have anything to do with how fast you crossed the threshold, so long as you did. [Singularity.] The math professor named his observation after the center of a black hole, where the standard physical models of the early 1980s broke down and produced nonsensical answers. Note that it is the models that broke down, not reality itself. Smarter-than-human minds would presumably do something or other, the problem was predicting it in advance. The math professor said, a future that contains smarter-than-human minds is *different in kind*, in a way that you don't get just from soaring skyscrapers.

Unfortunately, our professor of mathematics, Dr. Vernor Vinge, was way too poetic in naming his idea. Vinge called his idea the Singularity, and it was such a beautiful term that everyone borrowed the living daylight out of it. They used it to talk about Moore's Law, which is a prediction, not an unprediction. They used it to talk about accelerating technological change, the rate at which gadgets shower down upon us, which is exactly what Vinge tried to say was relatively unimportant. Not many people remember the point Vinge originally tried to make, about the unpredictability of smarter-than-human intelligence.

When I first encountered Vinge's idea in its original form, it changed my life. Vinge's observation is probably one of the most important, most insightful points any futurist has ever made. And; in order to refute it; I need to first present my idea about how to quantify intelligence - how to measure the most powerful force in the universe.

## **Optimization**

[Measure?] What I shall actually try to quantify is not intelligence, but the work performed by intelligence. If you look at an airplane, there's all these pieces of machinery, wheels and gears, parts and components, that go into making it fly. A bird has feathers and hollow bones, a tail for steering, a metabolism that beats the wings fast enough to fly, all of that goes into the bird.

A human brain has hundreds of distinguishable maps and areas and modules. So we're not asking what the brain *is*, or even how it works, but what it *does*. The analogy for an airplane would be to define a concept of aerodynamic lift, apart from all those wheels and gears.

[Optimization.] So I now introduce the concept of an *optimization process*: a system which hits small targets in large search spaces to produce coherent effects.

[Airport.] Let's say that I'm visiting a distant city, and a local friend volunteers to drive me to the airport. I don't know the neighborhood. Each time my friend approaches a street intersection, I don't know whether my friend will turn left, turn right, or continue straight ahead. So I can't predict the individual turns. I can, however, predict the outcome of my friend's

unpredictable actions: we'll arrive at the airport. Even if my friend's house were located elsewhere in the city, and my friend made a completely different sequence of turns, I would still predict that we would arrive at the airport. Isn't this a strange situation to be in, from a scientific perspective? I can predict the *outcome* of a process, but I can't predict any of the intermediate steps.

[Calibrated.] Imagine that I'm playing chess against a smarter opponent, and I have a probability distribution over my opponent's moves. Like, I think there's a 10% probability he'll move here, a 5% probability he'll move there, and so on. This probability distribution might be well calibrated, in the sense that moves to which I assign a 10% probability happen about one time in ten. I can do this even if I'm not nearly as good a chess player, so long as I can make at least some kind of guess about which moves are totally stupid. In the worst case, if my opponent has fifty possible moves, I could assign each move a possibility of 2%, and that would automatically be well-calibrated. That's what's called a maximum entropy distribution, and it represents total ignorance of the opponent.

So I can assign a well-calibrated distribution to a chess opponent who's vastly better than me. It won't be an exact prediction, but it will be well-calibrated. [Randomized.] And I can imagine taking this probability distribution, and using it to make a randomized machine which plays with that distribution - a player who randomly makes moves at the exact frequency I assigned in my probability distribution over my smarter opponent. In both cases, I assign exactly the same probability that my opponent will make a particular move. But in one case, I expect to lose the game to a superior opponent, and in the other case, I expect to crush an opponent that will sometimes randomly make bad moves. Even though it's exactly the same probability distribution! When I see the randomized player make a move that I assigned a very low probability, I chuckle and rub my hands, because I think the other player has randomly made a very poor move and now I can sweep him off the board. When a superior opponent surprises me by making a move to which I assigned a tiny probability, I'll groan because I think the other player has seen something I didn't see, and now I'm about to be swept off the board. And the moral here is that the creative unpredictability of intelligence is not at all like the entropic unpredictability of a random number generation. You can be exactly as uncertain about the actions, assign exactly the same probability distribution, and yet draw very different conclusions about the eventual outcome.

[Creative.] The lesson is that there's no such thing as a creative surprise without some identifiable criterion that makes it surprisingly effective. If there is no criterion by which you can identify the action as surprisingly effective, then you don't have intelligence, you just have noise. When I play chess against a smarter player, I can't predict *exactly* where my opponent will move against me. If I could predict that, I would be at least that good at chess myself. But I can predict the end result of my opponent's moves, which is a win for the other player. The more I'm surprised, the more confident I become of this outcome. The unpredictability of intelligence is a very special and unusual kind of surprise, which is not at all like entropy or randomness. There is a bizarre balance between the unpredictability of actions and the predictability of outcomes.

[Vinge.] This is my objection to Vernor Vinge's concept of the Singularity. It's possible that there would be something we could predict about the outcome after the Singularity, providing that we knew something about the goals of the transhuman intelligence - if we knew the criterion under which the transhuman's surprising actions were surprisingly effective. [Helpful.] That's not just a minor quibble. If we knew we were dealing with a transhuman whose actions were surprisingly helpful, we could predict that we would end up in a surprisingly pleasant future, as opposed to a surprisingly unpleasant future. The unpredictability of intelligence isn't the same as noise.

I will return to this point later. For now, back to optimization processes.

[Quantify.] We are now ready to propose a way of quantifying optimization, and therefore quantifying the work done by intelligence.

Back to the problem of quantifying intelligence. Consider a car - say, a Toyota Corolla. The Toyota Corolla is made up of some number of atoms, say, on the rough order of ten to the thirtieth. If you consider all the possible ways we could arrange those atoms, it's clear that only an infinitesimally tiny fraction of possible configurations would qualify as a useful working car. [Corolla.] A tiny fraction of the configuration space does describe vehicles that we would recognize as faster, more efficient, and safer than the Corolla. The Corolla is not optimal. But the Corolla is highly optimized, because the designer had to hit an infinitesimally tiny target in design space just to create a working car, let alone a car of the Corolla's quality. If you imagine all the ways that you could configure the atoms in the Corolla's air

conditioner that would mess up the air conditioning without messing up the whole car, it's clear that even within the space of working cars, most cars are not as good as the Corolla. The better the car you want, the more optimization pressure you have to exert. You need to exert substantial optimization pressure just to get a car at all. You can't build so much as an effective wooden wagon by sawing boards randomly and nailing them according to coinflips. To hit such a tiny target in configuration space, you need a powerful optimization process.

All of this discussion implicitly assumes that the designer of the Corolla was trying to produce a "vehicle", a means of travel. This proposition should be stated explicitly, but it is, in fact, correct, and it is highly useful in understanding the Corolla. The Corolla is a creative design because it is a surprisingly effective vehicle, relative to what you would get if you randomly reconfigured the atoms.

[How small?] Planning also involves hitting tiny targets in a huge search space. On a 19-by-19 Go board, there are roughly 10 to the 180 legal positions. In early positions of a Go game, there are more than 300 legal moves per turn. The search space explodes, and most moves are foolish moves if your goal is winning the game. From all the vast space of Go possibilities, the players seek out the infinitesimal fraction of plans which have a decent chance of winning.

[Bits.] This leads into an obvious way of quantifying optimization, relative to some assumed preference ordering over outcomes. We measure optimization in bits, not in programmer's bits like data on a hard drive, but in mathematician's bits that let you make yes-or-no choices. To figure out how much information is in the Toyota Corolla, we don't look at how many bytes it takes to store the design plans. Instead we ask how small is the infinitesimal fraction which measures how many possible configurations of atoms are vehicles just as good or better than the Toyota Corolla. We take the total volume of possibilities that are equal to or better than the Toyota Corolla under our preferences, divide that into the entire volume of possibilities, and take the logarithm. If you can hit a target volume of outcomes that's twice as small, you have one more bit of optimization power. Remember, the more you like an outcome, the fewer other possible outcomes will be as good or better. So if you want a really good outcome, you have to hit a really small target and exert a lot of optimization power to get there. That's my suggestion for quantifying optimization.

[Optimizers.] Human intelligence is one kind of powerful optimization process, capable of winning a game of Go, or refining silicon sand into digital computers - both very narrow targets to hit. Another optimization process is natural selection, which can hit the narrow pattern of a functioning butterfly in the huge space of possible genomes. Natural selection is much slower than human intelligence, but over geological time, *cumulative* selection pressure qualifies as a powerful optimization process.

## Evolution

Compared to the human brain, natural selection is a far simpler optimization process. Evolution is so simple that we can use simple math to describe how much work it does. [Fixation.] Suppose there's a gene which has a 3% fitness advantage relative to its alternatives at that allele; in other words, an individual with this gene has, on average, around 3% more children than others. Imagine that one mutant is born with this advantageous gene. What is the probability that this gene spreads through the whole population, as opposed to dying out? This calculation turns out to be independent of most things you would expect it to depend on, like population size, and the answer turns out to be six percent. The general rule is that if the fitness advantage is  $s$ , then the probability of fixation is  $2s$ . So even if you have a mutation that confers a 3% advantage in fitness, which is huge as mutations go, the chance is only 6% that the mutation spreads to the whole population. [Mean time.] Suppose the beneficial mutation does spread. How long does it take to become universal in the gene pool? This calculation does depend on population size. With a fitness advantage of 3%, and a population size of 100,000, the mean time to fixation is 767 generations.

For humans that would mean at least *ten thousand years* to accumulate a *single* beneficial mutation.

[Speed limit.] There's also a bound on the total amount of information evolution can produce in each generation, which works something like this: if two parents have sixteen children and two children survive to maturity and reproduce, then that's at most three bits of selection pressure per generation, to be divided up among all the genes currently selected on. That's not actually how you derive the bound, but it's a heuristic illustration of how the bound works.

In the real world, evolution is noisy - it does not perfectly eliminate all those and only those individuals with disadvantageous genes. Among mammals, you can probably go by the rule that each generation adds at most one bit of information to the gene pool. Again, this isn't bits like data on a hard drive, it's bits of information in gene frequencies. If you look at human DNA, you see three billion bases, and each base is selected from four alternatives, so that's two bits of data per base - six billion bits like bits on a hard drive. [Complexity wall.] But human DNA can't possibly contain six billion bits of useful information. The mutation rate of mammalian DNA is around  $10^{-8}$  per base per generation. Each DNA base has one chance in a hundred million of mutating in each offspring. That's a degenerative pressure, since the vast majority of mutations are negative. If the environment doesn't select out the negative mutations, the information will be lost. If you have around one bit of selection pressure per generation, and bases mutate at the rate of around one mutation per hundred million bases, then you can support around a hundred million bits of information against degeneracy pressures of natural mutation. The human genome probably contains substantially less information than this; there are around 30,000 actual protein-encoding genes, though regulatory regions add something to that.

Genomes on this planet today are probably not significantly more complicated than genomes five hundred million years ago. Life has been around on this planet for 3.85 billion years, and it only takes a hundred million generations to hit the complexity wall on how much information DNA can support. After that, once you gain one chunk of complexity, you lose something else - you've only got so much selection pressure to go around and maintain the information. Humans are smarter than chimpanzees but our muscles have become much weaker.

[Incredibly slow.] So we can calculate how strong natural selection is, and in any given period of time, it's extraordinarily weak. The only reason natural selection can produce patterns as complex as living beings is that, over the course of hundreds of millions of years, you get strong cumulative selection pressures - powerful enough to hit the target of a rabbit genome, in all the space of possibilities. But there's still an upper bound on what evolution can do.

## **Foundations of Order**

[Sources.] Is there any other way you can get a complex pattern besides evolution and intelligence? How about emergence - can complex patterns emerge? Stars can have very intricate internal dynamics, sunspots and corona and flares. But there's no use looking inside a star for optimization or complex functional design - there are no properties of the star that are there *because* they help the star burn brighter, or last longer. A star is an accident, an emergent thing; something that just happens when interstellar dust comes together. Before mind, before life, the universe was populated by emergent patterns - by accidents. [Stable.] There was no one to look around, but if you had looked around at the universe before life, you would have seen patterns which occurred frequently, or patterns that lasted for a long time. A star is a stable pattern; once interstellar dust coalesces and ignites, it burns for a long time. A raindrop is a frequent pattern; individual raindrops don't last long, but they emerge again and again, an attractor in the chaos of clouds.

[Emergence.] When you look up at the night sky, you're more likely to see stars whose internal dynamics let them burn for a long time, and less likely to see stars whose internal dynamics make them quickly explode. In that sense, you're more likely to see patterns that fit a certain criterion. You're more likely to see surprisingly long-lived stars. But if by luck one star happens to burn unusually bright or unusually long, that doesn't make it any more likely that future stars will also last longer - its successful pattern isn't copied onto other stars. Emergence is not a powerful optimization process! It takes a very long time or a huge number of tries to produce a single lucky accident.

[Evolution.] Evolution builds on its successes and turns them into bigger and more complex successes. For this reason, natural selection produces complexity enormously faster than emergence. I would bet there is less complexity in the entire Milky Way outside Earth than there is in a single butterfly.

[Foundations.] In this room there are three kinds of patterns. There are emergent patterns, like the flow of air currents. There are evolved patterns, like you and I and dust mites. And there are designed patterns, like the microphone I'm speaking into, and the clothes I'm wearing, and the room I'm standing in. Emergence, evolution, and intelligence. The three sources of pattern; the three foundations of order.

[Slow.] Just as evolution produces complexity enormously faster than emergence, so too does intelligence work enormously faster than evolution.

One of my pet peeves is when people talk about emergence, evolution, and intelligence all in the same breath, as if they were all the same thing. People read about emergence as if it were this blankly mysterious force that does nice things, and evolution were a blankly mysterious force that does nice things, and intelligence were a blankly mysterious force that does nice things, so why not confuse all of them together? Someone says, technology evolves. Technology does not evolve, it is designed. There's a *big* difference between those two foundations of order. Someone says that the order of an ant colony emerges from the ants. That order is the result of strong selection pressures on ant colonies, it didn't just pop out of nowhere. You may have heard people speculating that the Internet will just wake up, emergently, as the result of spontaneous order in TCP/IP packets routing around. I think Pat Cadigan put it best when she compared the idea of the routing fabric spontaneously waking up to putting dirty shirts and straw in a wooden box for the spontaneous generation of mice. Intelligence requires complex order; it can evolve, but it can't possibly emerge. People talk about corporations reproducing and dying and say, corporations are evolving. But if so, corporations are at most a dozen generations into the RNA world, because you have to run natural selection for literally millions of generations before it produces complex order.

I spoke earlier on how it takes dozens of tries and thousands of years for a single beneficial mutation to rise to fixation in the gene pool. For evolution to produce complex machinery, with multiple pieces that depend on one another, the pieces have to evolve one after the other, serially, sequentially. Natural selection has no foresight, it doesn't look ahead. If gene B depends on gene A to work, then gene A has to be reliably present before gene B presents a selective advantage. Evolution requires literally millions of years to produce complex adaptations, complex functions. A good human programmer can create a computer program with hundreds of interdependent parts in a *single afternoon*.

Relative to natural selection, there are no words to describe just how unprecedented human intelligence really was. There was no possible way you could have imagined the power of intelligence by looking at what came before. Maybe you would have imagined a magical optimization process that would create complex order in mere millennia, instead of eons; but to do

it in a single afternoon? The rise of human intelligence was a breaking of induction, a black swan in the foundations of order.

[Humans slow.] And yet human intelligence doesn't seem anywhere near the top of the line. Our neurons run at the absurdly slow speed of 100 cycles per second; computational neurobiologists use the hundred-step rule, which is that any neural algorithm has to work in a hundred serial steps or less. We have sensory modalities, like vision and sound, which are great for playing baseball, but not so good for algebra or computer programming. We're born with a limited amount of computing hardware, and we can never get any more; instead of adding new capacity as we grow, our brains slowly disintegrate as we age. Above all, we have no direct access to our own neurons; we can't rewrite the brain's neural circuitry to make ourselves smarter. We can't self-improve.

Let us consider again this notion of three eras, of emergence, evolution, and intelligence. [When end?] When would we mark the end of the reign of emergence and the beginning of the reign of evolution? Where would we locate the boundary between survival of the stable and replication of the fittest? Should we look to the first replicator, the ultimate grandparent of all life? [Replicator.] That's the usual answer. But the first replicator was, itself, an accident - an emergent pattern - a product of the era of stable things. Otherwise evolution could never have begun. From our later perspective, that first replicator ever to exist must have looked like one of the oddest and most awkward sights in the universe - a replicator with the pattern of a stable thing, a reproducing structure with no evolutionary optimization whatsoever. An onlooker might mistake that first replicator for an exceptionally stable thing, a pattern that emerged often from primordial soup. A great success, to be sure, but nothing *truly* different. The true nature of the next era would not come clear until you saw the second replicator ever to exist, your *penultimate* grandparent, an optimized pattern that would never emerge from primordial soup without a billion copies of that first replicator to mutate. It was the second replicator that was a new kind of structure, a pattern you would never find in a world of stable things alone.

[Evolved human.] And what a strange thing a human must be - an intelligence inscribed entirely by evolution, a mind that wasn't constructed by another mind. We are one of the oddest and most awkward sights in the universe - an intelligence with the pattern of an evolved thing. We only had

to evolve once, but we did have to evolve. The changeover from evolution to intelligence, from replication of the fittest to recursive self-improvement, could never have happened otherwise.

[Mind.]

Think of that, when you hear the word "Artificial" in the phrase "Artificial Intelligence". It's a rather unpoetic way, one would think, to describe the first mind born of mind.

## **RSI**

[RSI.]

The significance of Artificial Intelligence is that it closes the loop between intelligence and technology - intelligence creating technology that improves intelligence, a positive feedback cycle. The AI rewrites itself and makes itself smarter, which increases the AI's ability to rewrite itself, which allows the AI to rewrite itself yet again and make itself even smarter. [Good.] The idea of recursive self-improvement is usually attributed to a Bayesian theorist named I. J. Good - it's odd how many of the people involved with this are math professors. I. J. Good called the hypothetical positive feedback effect an "intelligence explosion".

To a limited extent, human beings improve themselves; we learn, we practice, we hone skills and knowledge. To a limited extent, these self-improvements can even increase our ability to improve ourselves further. New discoveries can increase our ability to make further discoveries - in that sense, knowledge feeds on itself. [Weak.] But there is still an underlying level that we can't touch, a protected layer, the underlying brain architecture. We can't, for example, give ourselves a sensory modality for algebra or source code - we're stuck using the visual cortex to visualize everything. Our brains are ultimately the source of discovery, and our brains are much the same today as ten thousand years ago. Similarly, in evolution, beneficial mutations can open up the way for new beneficial mutations; in this sense, adaptation feeds on itself. But natural selection doesn't rewrite the process of natural selection; new adaptations don't change the nature of evolution, the lack of abstract intelligence, the lack of foresight, the reliance on random mutations, the blindness and incrementalism. Likewise, not even the invention of science changed the essential character of the human brain:

the limbic core, the cerebral cortex, the prefrontal self-models, the characteristic speed of 100Hz.

[Strong:] A mind that can understand itself and understand its own nature as an optimization process, a mind that can reach down and directly rewrite the layer that does the heavy work of optimization - that mind wraps around far more *strongly* than either evolution producing adaptations in the genome, or a human brain acquiring content and knowledge.

Surface similarities can't make up for huge quantitative differences. Is human intelligence nothing novel relative to evolution, just because both human intelligence and evolution can produce complex, functioning machinery? But evolution does one bit per year per species and produces complex machinery over millions of years, while humans do one bit per second per human and produce new complex machinery in an afternoon. Surface similarity or not, the quantitative difference is so huge as to count as a qualitative difference in practice. Humans writing books about the philosophy of science may technically be thinking about how to think. But I suspect we'll see a huge practical difference between a mind that can ponder philosophy of science and a mind that can directly rewrite its own source code.

[Slow.] I sometimes encounter skepticism about I. J. Good's intelligence explosion, because progress in AI has the reputation of being very slow. Consider a loose historical analogy. [Rutherford.] In 1933, Lord Ernest Rutherford said, quote, "Anyone who looked for a source of power in the transformation of atoms was talking moonshine." Rutherford was an experimental physicist, so he was probably thinking about the tremendous difficulties that he had to go through to fission a tiny handful of nuclei in his laboratory.

[Chain.] In a fission chain reaction, the key number is  $k$ , the effective neutron multiplication factor.  $k$  is the average number of neutrons from a fission reaction that cause another fission reaction. When Enrico Fermi and a team of physicists built the first nuclear pile in 1942, in a squash court beneath Stagg Field at the University of Chicago, the first-ever critical reaction had  $k$  of 1.0006. The pile was made of alternate layers of graphite and uranium bricks, and Fermi had calculated that the pile would reach criticality between layers 56 and 57. Cadmium absorbs neutrons, so the builders used wooden rods covered with cadmium foil to keep the pile from

reacting while they were building it. Layer 57 was completed on the night of December 1st, 1942. On December 2nd, 1942, they started pulling out the cadmium control rods. They would pull out the rod a few inches, then another few inches, then another few inches, and each time they pulled it out, the reaction would increase. There was a pen whose trace showed the neutron intensity, and each time they withdrew the rod more, the trace would climb, and then level out. At 3:25 pm, Fermi ordered the last control rod withdrawn another 12 inches, and Fermi said, "This is going to do it. Now it will become self-sustaining. The trace will climb and continue to climb. It will not level out." And the clicks of the geiger counter came rapidly, and more rapidly, and merged into a roar. The pen on the graph recorder climbed and continued to climb. Fermi kept the pile running for twenty-eight minutes, and the neutron intensity doubled every two minutes.

That was the first nuclear reaction, with  $k$  of 1.0006. [Prompt.] The only reason the pile was controllable was that some of the neutrons from a fission reaction are delayed; they come from the decay of short-lived fission byproducts. A nuclear reaction that is prompt critical is critical without the contribution of delayed neutrons. If the pile had been prompt critical with  $k$  of 1.0006, the neutron intensity would have doubled every *tenth* of a second, instead of every two minutes. Nuclear weapons are prompt critical and have  $k$  in the range of two or three.

[Lessons.] So there's a couple of morals I want to draw. One moral is that there's a qualitative difference between one self-improvement triggering 0.9994 further self-improvements, and one self-improvement triggering 1.0006 further self-improvements. Another moral is not to confuse the speed of AI *research*, done by humans, with the speed of a real AI once you build it. That is mixing up the map with the territory, baking the recipe into the cake. It took years for a small group of physicists to achieve the understanding and the theory, the skill and the methods, that let them build the first nuclear pile. But, once the pile was built, interesting things happened on the timescale of nuclear interactions, not the timescale of scientific research. In the nuclear domain, elementary interactions happen much faster than human neurons fire. Much the same may be said of transistors.

Of course I do not think there is any direct mathematical analogy between an intelligence explosion and a nuclear explosion. For one thing, nuclear weapons don't intelligently seek out ways to increase their own neutron

multiplication factor. So it's quite possible that it will take years for a group of AI theorists to understand how to build an AI, and then they build the AI, and it goes FOOM. Thus we would see, for the first time, the true shape of the era of intelligence; the power of a mind born of mind. [True shape.]

## **Anthropomorphism**

Then what happens?

Before I can discuss that question, I need to deliver some dire warnings. [Dire.] Norman R. F. Maier said: "Do not propose solutions until the problem has been discussed as thoroughly as possible without suggesting any." There is experimental evidence backing this up; studies of group problem-solving where groups who were told to hold off on proposing a solution until the end of the session proposed much better solutions than groups who were not told this. Robyn Dawes gives us a further caution, that the tougher the problem you face, the more likely people are to propose solutions immediately. Some of the problems I want to talk about are so incredibly difficult that it is not possible to talk about them at all, because people solve them instantly.

There is a fundamental reason why humans would answer questions about Artificial Intelligence much too quickly; we have a lifetime of experience and even built-in brain mechanisms that will give us the wrong answers.

[Human universals.] The anthropologist Donald Brown once compiled a list of more than 200 "human universals", a few of which I've shown here. These characteristics appear in every known human culture, from Palo Alto to Yanomamo hunter-gatherers in the Amazon rain forest. They're characteristics that anthropologists don't even think to report, because, like air, they're everywhere. Of course the reports that make it into the media are all about differences between cultures - you won't read an excited article about a newly discovered tribe: They eat food! They breathe air! They feel joy, and sorrow! Their parents love their children! They use tools! They tell each other stories! We forget how alike we are, under the skin, living in a world that reminds us only of our differences.

Why is there such a thing as human nature? Why are there such things as human universals? Human universals aren't truly universal. A rock feels no pain. An amoeba doesn't love its children. Mice don't make tools.

Chimpanzees don't hand down traditional stories.

[Complex universal.] Complex biological machinery requires more than one gene. If gene B depends on gene A to produce its effect, then gene A has to become nearly universal in the gene pool before there's a substantial selection pressure in favor of gene B. A fur coat isn't an evolutionary advantage unless the environment reliably throws cold weather at you. Well, other genes are also part of the environment. If gene B depends on gene A, then gene B isn't a significant advantage unless gene A is reliably part of the genetic environment.

Let's say that you have a complex adaptation with six interdependent parts, and that each of the six genes is independently at ten percent frequency in the population. The chance of assembling a whole working adaptation is literally a million to one. In comic books, you find "mutants" who, all in one jump, as the result of one mutation, have the ability to throw lightning bolts. When you consider the biochemistry needed to produce electricity, and the biochemical adaptations needed to keep yourself from being hurt by electricity, and the brain circuitry needed to control electricity finely enough to throw lightning bolts, this is not going to happen as the result of one mutation. So much for the X-Men. That's not how evolution works. Eventually you get electric eels, but not all at once. Evolution climbs a long incremental pathway to complex machinery - one piece at a time, because each piece has to become universal before dependent pieces evolve.

[Psychic unity of humankind.] When you apply this observation to the human mind, it gives rise to a rule that evolutionary psychologists have named *the psychic unity of humankind*. Any piece of *complex* machinery that exists in the human mind has to be a human universal. In every known culture, humans experience joy, sadness, disgust, anger, fear, and surprise. In every known culture, humans indicate these emotions using the same facial expressions. The psychic unity of humankind is both explained and required by the mechanics of evolutionary biology.

When something is universal enough in our everyday lives, we take it for granted; we assume it without thought, without deliberation. We don't ask whether it will be there - we just act as if it will be. In the movie *The Matrix*, there's a *so-called* Artificial Intelligence [Agent Smith] named Smith... Agent Smith. At first Agent Smith is cool, dispassionate, emotionless, as he interrogates Neo. Under sufficient emotional stress,

however, [Morpheus] Agent Smith's cool breaks down. He vents his disgust with humanity and, yes, lo and behold, his face shows the human-universal expression for disgust.

Back in the era of pulp science fiction, magazine covers occasionally showed a bug-eyed monster, colloquially known as a BEM, carrying off an attractive human female in a torn dress. [BEM.] For some odd reason, it's never an attractive man in a torn shirt. Don't blame me, I didn't draw it.

Would a non-humanoid alien, with a completely different evolutionary history, sexually desire a human female? It seems rather unlikely. People don't make mistakes like that by explicitly, deliberately reasoning: "All minds are likely to be wired pretty much the same way, so presumably a bug-eyed monster will find human females attractive." Probably the artist did not even think to ask whether an alien *perceives* human females as attractive. Instead, a human female in a torn dress *is sexy* - inherently so, as an intrinsic property. They who made the mistake did not think about the evolutionary history and cognitive mechanisms of the bug-eyed monster. Rather they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the bug-eyed monster doesn't enter into it.

[Projection.] This is a case of what E. T. Jaynes called the "mind projection fallacy", which happens when you project mental properties onto the outside world. 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.

Returning to anthropomorphism, the lesson is that you don't even need to realize you are anthropomorphizing in order to do it. You can get back a wrong answer without even realizing that there exists a question. [BEM.] Since this woman is obviously very *attractive*, of course the monster is *attracted* to her - isn't that logical? When you reason about other minds which are not human, every step of the process can be contaminated by assumptions you take no more notice of than oxygen. You hose yourself the moment you put yourself into the shoes of a nonhuman optimization process. Biologists before the 1960s used to do this with natural selection, and did they ever get in trouble! They would expect evolution to do things that humans would do if humans were designing organisms, and needless to say their predictions went astray. In 1966 a biologist named George Williams wrote a very famous book called "Adaptation and Natural Selection" which tore apart some of the grosser fallacies, and today things

are saner. But the moral is that your ancestors evolved in an environment where every powerful intelligence they met was another human being. So your brain has evolved to model other minds through empathy; asking what you would do in the other mind's shoes. This is going to give you wrong answers if the other mind is not, in fact, human.

## **Minds in general**

[Anthrop.] So if you can't predict what an AI will do through anthropomorphism, if you can't trust your instincts that evolved to model other humans, how are you supposed to predict the AI? How can we guess what AIs will do?

[Trick.] Actually, this is a trick question. Here's why.

[ATP.] Evolution strongly conserves some structures - once other genes evolve which depend on a previously existing gene, that early gene is set in concrete; it can't mutate without breaking other adaptations. ATP synthase is a molecular machine which is essentially the same in animal mitochondria, plant chloroplasts, and bacteria - it 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.

[Mindspace.] Asking what "AIs" will do is a trick question because it implies that all AIs form a natural class. Humans do form a natural class because we all share the same brain architecture. But when you say "Artificial Intelligence", you are referring to a vastly larger *space of possibilities* than when you say "human". 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; 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.

You should resist the temptation to generalize over all of mind design space. If mind design space contains two to the trillionth power minds whose makeup can be specified in a trillion bits or less, then every universal generalization that you make has two to the trillionth power chances to be

falsified. The *only* reason you could find yourself thinking that you know what a fully generic mind will do, is if you put yourself in that mind's shoes - imagine what you would do in that mind's place - and get back a generally wrong, anthropomorphic answer. Now if I said there was *no* mind that did the same thing you would, that would also be a universal generalization, which wouldn't just be easy to falsify, it's already falsified. By you. *You* do the same thing you would in your shoes. Some other minds may do that to. Others won't. You can't talk about AIs in general - you have to specify something about which mind-in-general you're talking about, say something about the architecture or the cognitive content. There's a lot more room in the category "AI" than there is in the category "human". So it is nonsense to ask what an AI will do. You have to specify what kind of AI.

### **Giant Cheesecake Fallacy**

One often hears, in futurism, a line of reasoning that goes something like this. [Fallacy.] Someone says: "When technology advances far enough, we'll be able to build minds far surpassing human intelligence. Now it's clear, that if you're baking a cheesecake, how large a cheesecake you can bake depends on your intelligence. A superintelligence could build enormous cheesecakes - cheesecakes the size of cities. And Moore's Law keeps dropping the cost of computing power. By golly, the future will be full of giant cheesecakes!" I call this the Giant Cheesecake Fallacy. [Giant.] It happens whenever someone leaps directly from *capability* to *actuality*, without considering the necessary intermediate of *motive*.

[Missing.] Here are some frequently heard lines of reasoning that include a Giant Cheesecake Fallacy:

- A sufficiently powerful Artificial Intelligence could overwhelm any human resistance and wipe out humanity. (Whisper: And the AI would decide to do so.) Therefore we should not build AI. [Cheesecake.]
- A sufficiently powerful AI could develop new medical technologies capable of saving millions of human lives. (Whisper: And the AI would decide to do so.) Therefore we should build AI.
- Once computers become cheap enough, AIs will be able to perform the vast majority of jobs more easily than humans. (Whisper: 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.

The most important property of any optimization process is its target - the region into which the optimizer tries to steer the future, the criterion which the patterns it produces will fit surprisingly well. That's what determines the effect of the optimizer on the real world. If you're talking about a smarter-than-human AI, then you're talking about a really huge impact on the world, an extremely powerful force steering the future. So the target of the AI, the specific region into which the AI tries to steer the future, becomes extremely important.

### **Friendly AI**

[Minds in general.] The temptation is to try to reason about the possibilities by asking what "AIs" will "want", forgetting that the space of minds-in-general is much wider than the tiny human dot. Resist the temptation to generalize over all possible minds! That's a storyteller's habit. Storytellers spin tales of the distant and exotic land called Future in much the same way they might tell stories of Atlantis. And they talk about AIs as if AIs were some pre-existing race inhabiting Atlantis. Storytellers, to entertain their audiences, spin tales of how things *will be*; they make *predictions*.

[Engineers.] When civil engineers build a bridge, they don't spin stories of how bridges in general will stay up. Civil engineers don't even need to be able to say, of an *arbitrary* bridge design, whether it will fall or stay up. Civil engineers only need to *choose one particular* bridge design, of which they *can* say that it supports at least 30 tons. In computer science, there's a theorem, Rice's Theorem, which generalizes the halting theorem. [Rice.] Rice's Theorem states that it is not possible to distinguish whether an *arbitrary* computer program implements *any* function, including, say, simple multiplication. And yet, despite Rice's Theorem, modern chip engineers create computer chips that implement multiplication. Chip engineers select, from all the vast space of possibilities, *only* those chip designs which they *can* understand.

And so now we are ready to pose the problem. Again, be careful not to try to solve this problem immediately; think about it as thoroughly as possible, trying to figure out what all the dimensions are, before you allow any

thought of a solution to enter your mind. It's hard to even discuss this problem with people, since it's so incredibly difficult that most people solve it instantly. [Friendly] We need to build an AI such that its optimization target, the region into which the AI tries to steer the future, is, knowably, friendly, nice, good, helpful... At this stage of discussing the problem, we aren't ready to be any more specific than that, except to say, for example, that if we build an AI that transforms everything in its reach into cheesecake, we probably could have done better than that. And, we need the AI's impact on the world to still be positive, even if the AI is smarter than human. This rules out strategies like pointing a gun at the AI and telling it to behave or else, unless you think that intelligence is no match for a gun. And, this property of the AI, its friendliness, has to be stable even if the AI is modifying its own source code and recursively self-improving.

[Kurzweil.] Ray Kurzweil, who devoted a chapter of *The Singularity Is Near* to the perils of advanced technology, had this to say about the perils of AI: "There is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence." This objection usually pops up, in one form or another, when I talk about creating a Friendly AI. Usually it's put as follows: If the Friendly AI can modify its own source code, it will be able to modify itself to no longer be Friendly. [Cheese.] What you should spot here is a Giant Cheesecake Fallacy; just because the AI has the *capability* to modify itself in a particular way doesn't mean it has the *motive*.

[Gandhi.] Right now, I am friendly in some intuitive senses of the term; for example, I assign a large negative utility to killing people. I could, if I wished, adopt other ethical principles; I could try to make myself enjoy killing people. But I don't want to do that, because currently, I prefer that people not die. If I could modify my own source code, I would not modify myself to enjoy killing people.

Humans come from the womb with a whole complex set of drives programmed in by evolution. If you want a human to do something *against* their drives, you have to persuade or bribe or threaten them into it. With an AI, you're standing in the place of natural selection, choosing the drives. The programmer stands not just *in loco parentis*, but *in loco evolutionis*. You are not trying to externally constrain the AI to be Friendly against its own desires; you are trying to create an AI that is Friendly by its own internal nature. If the AI starts *wanting* to modify itself to no longer be

Friendly, it has *already* stopped being Friendly at that point.

More generally, we can say that most utility functions will be trivially consistent under reflection in expected utility maximizers - if you want to accomplish something, you will want to keep wanting to accomplish it, because you know that if you stop wanting to accomplish it, it's less likely to get accomplished.

This seems intuitively obvious, but I am presently not able to deliver a mathematical proof of it. [Proof.] The reason is that current decision theory revolves around a formalism called expected utility maximization, and the formalism can't handle self-reference. You can use expected utility to choose between actions, or choose between different versions of source code that choose between actions. But there's no way to wrap all away around - no way to modify the source code that does the modifying. The theory breaks down on an infinite recursion. The problem is not representable within the system - you can't solve it even given infinite computing power. And yet humans think about themselves all the time; we have no difficulty using the word "I" in a sentence. I can think about what it would be like to modify my own neural circuitry, and my brain doesn't crash. How are humans handling the self-embedding? The significance of this problem is not that the classical algorithm is too inefficient for the real world, but that we don't understand the structure of the problem; we don't know what kind of work to perform.

Now let's pull back and talk about the significance of the problem. [Significance.] Intelligence is by far the most powerful of the foundations of order, capable of creating complex machinery in a single afternoon, instead of millions of years. Intelligence supports everything we do as humans, from math to persuasion, from chess to music; it's the root of all our technology from MP3 players to fire, and also every weapon from guns to nukes, everything that makes a human a more dangerous opponent than a chimpanzee. The rise of human intelligence has significantly transformed the world, we've carved our faces into mountains and set footprints on the moon. A smarter-than-human AI potentially has an impact on the world larger than everything we've done up to this point. If an intelligence explosion occurs, and by the way, I think it will, the impact is even huger, it just blows the imagination away completely. That is a *heck* of a lot of power. And the most important property of that power will be the region into which it steers the future.

[Precious.] The most important property of ourselves, of our own intelligence as humans, is our treasuring of the things we treasure: A sense of fun, the love of beauty, taking joy in helping others, the ability to be swayed by moral argument, the wish to be better people. It is astounding that conscience and compassion should emerge from natural selection, from billions of years of lethal competition and combat. And yet we have conscience. We have compassion. How did these things evolve? That's a real question with a real answer, which you can find in the field of evolutionary psychology. But the point is that love did evolve, and now we have to keep love alive. The light in us is the most important gift we have to give to the future, because if minds stop wanting to do these things, they won't get done. In the Milky Way there are four hundred billion stars. In the visible universe there are sixty billion galaxies. All waiting to be shaped by intelligence. If we do our jobs right, and *only* if we do our jobs right, then someday, four billion years from now, among all those stars and galaxies, there will still be told that story of how love first came into existence among *Homo sapiens*, and was preserved, from parent to child in unbroken chain, as minds that love design more loving minds. That is our responsibility, we who are the first intelligences ever to exist, we who are the link pin connecting the future and the past; we must choose, by our own will and intelligence, the shape of the next era, and shape it to light.

[Cheese.] And the scary part is this: We have to solve that problem in the real world. It is not a morality tale with a predetermined happy ending. We have to fulfill that responsibility using only human intelligence, our tiny little brains at the uttermost dawn of mind, as awkward as that first replicator. We have to face that challenge in the real world where Murphy's Law reigns supreme. [Cake.] Here at the dawn where intelligence begins, the chain of light is most fragile, most easily broken.

[Intelligence.] We have to shape the next era using our own human intelligence, and for intelligence to be useful, you have to actually use it. If you say, "Oh, what a wonderful story," and then you smile happily and go back to your daily job with an uplifted spirit, and you don't actually *do* anything about it, that is *not* invoking intelligence to solve the problem. In this world there are highly paid, well-trained professionals whose job it is to sell lipstick, because the owners of the lipstick corporation care too much about selling lipstick to leave it up to unpaid volunteers. We should give at least that much consideration to the next billion years. The rules don't

change, no matter the stakes, whether it is a question of invoking human intelligence to cure cancer, or a question of invoking human intelligence to shape the next billion years. You need people working full-time on that specific problem.

[Friendly] I've only touched on the barest beginnings of the challenge of Friendly AI. There are many subproblems here, and all sorts of fascinating issues that you encounter once you start trying to work through exactly what would need to happen. Developing a rigorous, fully reflective decision theory - which is to say, understanding the type of work performed by a mind rewriting its own source code in a stable way - is part of it; it's one of the facets of the problem I'm working on right now. There's a lot more parts, but the important thing to remember is that it is a math problem. It is not fate, it is not destiny, it is a math problem. [Solve.] The Singularity Institute is presently seeking an exceptionally brilliant math talent who can think in code to help work out the mathematical foundations - you can see the job description online. The Singularity Institute doesn't have much funding, since this sort of thing is so huge, and so important, that everyone thinks someone else will take care of it. But if you're an exceptionally brilliant math talent who cares enough to work on the problem for free, the Singularity Institute may be able to pay you enough to live on so you can work on the problem full-time. And of course, if anyone here can help us with funding, that also would be greatly appreciated; this math problem doesn't receive a thousandth of the attention that humanity gives to buying and selling lipstick. But someone has to work on the problem, or it won't get done. When you want something, you have to invoke human intelligence to get it. It will take a specific effort, just like curing cancer or designing the first aircraft, to keep the light alive through the transition, to ensure that love does not perish from the universe. It is the most important math problem in the world.

Questions? Comments?